The movie database

Explications

L'exercice suivant est une analyse du jeu de données `tmdb_5000`. Ce jeu de données contient les informations de 5000 films ayant été produits ces dernières années.

Cet exercice sera commencé en classe commune et fera l'objet d'une note bonus pour l'examen final du cours *Python pour la datascience* pour 2020.

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

Vous aurez deux fichiers disponibles pour cette analyse :

- `tmdb_5000_movies.csv` ; ce fichier contient des informations pour 5000 films telles que : budget, genres, homepage, id, keywords, original_language, original_title, overview, popularity, production_companies, production_countries, release_date, revenue, runtime, spoken_languages, status, tagline, title, vote_average, vote_count
- `tmdb_5000_credits.csv` ; ce fichier qui peut être joint au fichier précédent contient les informations concernant les crédits du film en question.

Plus de détail concernant le jeu de données peuvent être trouvés ici.

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.

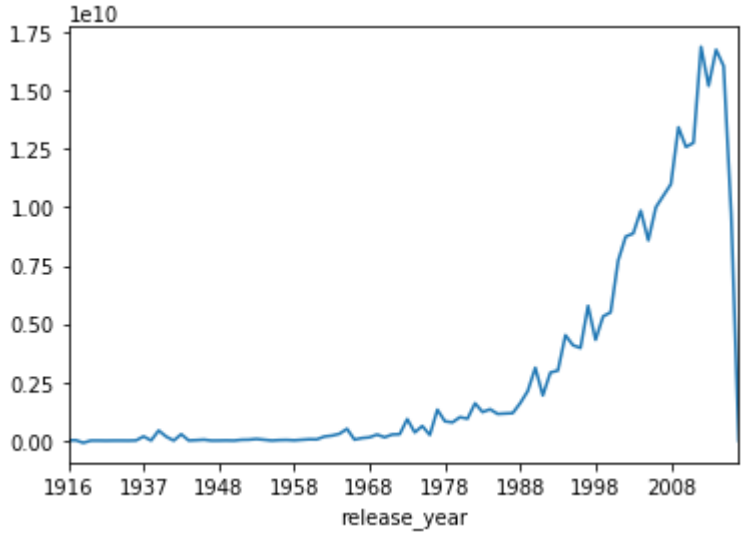
Nous ferons la simplification de langage de dire que l'industrie du cinéma est représentée par ces 5000 films. **De plus à chaque fois qu'une question sera posée, seront attendus la réponse à la question mais aussi le code pour obtenir la réponse.**

Questions

Q1. Charger les données dans deux DataFrame pandas nommées `credits` et `movies`.

Q2. Calculer le revenu total de l'industrie du cinéma sur ces 5000 films.

Q3. L'objectif ici est d'afficher un graphique représentant l'évolution du profit par année pour toute l'industrie. C'est à dire obtenir un graphique comme ci-dessous. Vous avez quelques questions permettant de décomposer le calcul pour obtenir ce graphique.



Q3.1. Créer une colonne `release_year` contenant l'année de sortie du film. Vous pouvez pour créer cette colonne supprimer les valeurs nulles.

Q3.2. Créer une colonne `profit` avec le profit de chaque film.

Q3.3. Créer le graphique précédent affichant le profit par année.

Q4. Quel est le film qui a été le plus profitable ? (Ici seront attendu la bonne réponse mais aussi la manière de trouver la bonne réponse)

Q5. Quel est le mot le plus présent dans la description des films ?

Q6. Afficher un diagramme en camembert des langues de sorties des films.

Q7. Quels sont les pays qui ont produits le plus de films ? Dans cette questions vous allez devoir travailler avec la colonne `production_countries` qui est sous le format json. Par exemple sur la première ligne vous trouverez dans une string :

```
'[
  {"iso_3166_1": "US", "name": "United States of America"},
  {"iso_3166_1": "GB", "name": "United Kingdom"}
]'
```

Il faudra lire cette information et la mettre au bon format pour la manipuler. Pour lire un format JSON vous pouvez utiliser le module JSON de la manière

suivante par exemple :

```
import json
j = "{...json stuff...}"
json.loads(j)
```

Q8. En vous aidant des indications de la question 7 trouvez quel est le genre le plus présent au cinéma.

Q9. Donner la note moyenne par genre (question de niveau élevé).

Q10. Dans combien de film a joué Brad Pitt ?

Q11. Quel était le premier film d'Angelina Jolie ?

Réponses

Vous repondrez 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 vendredi 3 avril 12h.

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