

TP 2 Exploration des Données

TP Data Science

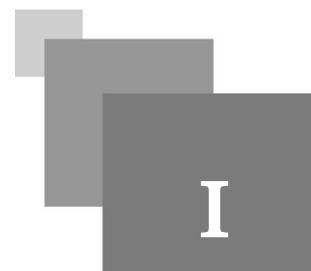
Ilyas Bambrik

Table des matières



I - Exercice : IMDB	3
II - Exercice : Rotten Tomatoes	5

Exercice : IMDB



Pour le DataSet *imdb_top_1000.csv*, lisez le fichier à partir de votre Jupyter Notebook et listez les 5 premières lignes afin de vérifier le contenu du DataSet.

Question 1

- Quel est le directeur le plus récurrent dans la liste des 1000 meilleurs films de IMDB ? Combien de fois ce directeur est apparu dans le DataSet?
- Quel est le meilleur film selon la colonne *IMDB_Rating* ?
- Listez les genres des films de le DataFrame. Quel est le genre le plus dominant dans le top 1000 ?
- Quel est le genre de films le plus rare dans le top 1000 ?

Question 2

- Quelle est l'intervalle des valeurs de la colonne *IMDB_Rating* (minimum, maximum) ?
- Listez les films qui ont obtenu le score minimal ?

Question 3

La colonne *Gross* indique les revenus monétaires du film mais celle-ci est exprimée comme une chaîne de caractères.

- Transformez la colonne afin d'avoir la colonne comme un nombre entier. Créez une colonne pour la série transformée nommée *Gross_int*.
- En utilisant le résultat de cette question, quel est le film qui a généré le plus gros succès commercial selon *Gross_int*.
- Quel est l'année de publication (*Released_Year*) avec le revenu total le plus large selon *Gross_int*.
- Créez un DataFrame contenant le nombre total de films pour chaque combinaison année de publication (*Released_Year*) et genre de film (*Genre*). Nommez la colonne indiquant le nombre films "*No of movies*".

Indice :

Vous devez éliminer les valeurs nulls et supprimer les caractères ',' de chaque valeur avec *map*.

Pour remplacer un (ou une séquence) caractère dans un string, vous devez utiliser la méthode *replace*.

Question 4

- Triez le DataSet selon la valeur de la colonne *Released_Year* d'une manière ascendante.
- Triez le DataSet selon *IMDB_Rating* descendante et *Runtime* ascendant.

Question 5

Meta_score est un score moyen accumulé depuis plusieurs sites de critique.

Question 6

- Affichez la description statistique de cette colonne.
- Transformez la colonne *Meta_score* afin que ces valeurs soient entre 0-9. Nommez le résultat dans le DataFrame : '*Meta_score_Transform*'
- Créez un nouveau score égal à la moyenne entre *Meta_score_Transform* et *IMDB_Rating*. Nommez le résultat comme une colonne du DataFrame : '*Final_Score*'
- Quel est le meilleur film selon '*Final_Score*' ?
- En observant le DataSet, pour quoi un classement selon le succès commercial n'as pas de sens ?

Question 7

Dans les colonnes Start1, Start2, Start3 et Start4, quel est l'acteur avec le nombre d'apparitions le plus élevé dans ce top 1000 ?

Exercice : Rotten Tomatoes

II

Pour le DataSet *rotten_tomatoes_movies.csv*, lisez le fichier à partir de votre Jupyter Notebook et listez les 5 premières lignes afin de vérifier le contenu du DataSet.

Question 1

- Listez les valeurs uniques de la colonne *content_rating*.
- En utilisant les colonnes *original_release_date* et *streaming_release_date*, quel est le nombre de films qui ont été directement publiés dans un service de streaming ?
- Quel est le nombre de films qui n'ont jamais été publiés en streaming ?

Question 2

- Listez les films qui ont obtenu le score minimal dans ce Dataset.
- Utilisez le résultat précédant afin de trouver l'entreprise de production (*production_company*) obtenant ce score minimal le plus fréquemment.

Question 3

- Pour chaque genre (colonne *genres*), trouvez le film avec le meilleur score selon la colonne *tomatometer_rating*.
- Pour chaque genre, trouvez le score minimal, maximal ainsi que la moyenne (utilisez la méthode *agg*).

Question 4

Quel est le nombre de films commun entre IMDP 1000 et ce DataSet de Rotten Tomatoes. Pour que votre réponse soit considérée comme juste, vous devez utiliser la jointure.