

TP 2 Exploration des Donnée

MI - IA / GL

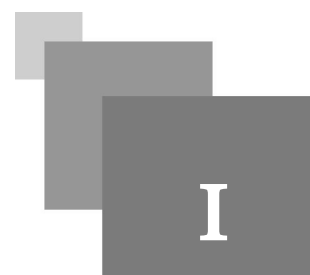
Ilyas Bambrik

Table des matières



I - Exercice : IMDB	3
II - Exercice : Sélection	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 se répète ?
- Quel est le meilleur film selon la colonne *IMDB_Rating* ?

Question 2

- Quelle est l'intervalle des valeurs de la colonne *IMDB_Rating* ?
- 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*.

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* et *Meta_score* descendante.

Question 5

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

Question 6

- Affichez la description statistique de cette colonne.
- Transformez cette colonne 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 ? Vous devez utiliser la méthode *apply* pour que votre solution soit considérée juste (il existe d'autre solutions évidemment).

Exercice : Sélection

II

La colonne *unaccredited* indique si l'université n'est pas reconnue pour des raisons réglementaires.

Question 1

- Quel est le type de cette colonne ?
- Trouvez les noms des universités non accréditées dans le classement s'ils existent.

Indice :

Utilisez l'attribut *dtype* de la colonne (Serie)

Question 2

- Combien d'universités Algériennes existent dans le classement ?
- Quelle est l'université Algérienne la mieux classée selon les données ?

Ce classement comporte plusieurs colonnes de score comme *scores_overall* et *scores_research*.

Question 3

- Sélectionnez les colonnes du DataFrame allons de la colonne *scores_overall* jusqu'à *scores_international_outlook_rank*.
- Sélectionnez les lignes du DataFrame avec une valeur *scores_research* ou *scores_teaching* supérieure à 80
- Sélectionnez les universités classées parmi les cinq meilleurs dans *selon scores_teaching_rank* ou *scores_research_rank* ou *scores_citations_rank*.