"Décoder le Silver Screen : aperçus du Dataverse IMDB"

Introduction :

Dans cet article, on partagera notre parcours d'utilisation de Python pour analyser les données imdb et découvrir des informations précieuses. On expliquera notre approche, présenterai nos découvertes et quelques idées intéressantes.

Notre travail est basé sur un fichier Excel qui s'appelle « imdb_dataset » qui se trouve sur la plateforme « Kaggle » qu'on va appliquer, avec logiciel Jupyter, sur lui les tâches suivantes :

1. Nettoyage des données.

2. L'analyse exploratoire des données.

3. Distribution des données et de l'échantillonnage.

Chapitre I : Nettoyage des données :

	title	type	release_year	age_certification	runtime	genres	production_countries	seasons	imdb_id	imdb
0	Five Came Back: The Reference Films	SHOW	1945	TV-MA	48	['documentation']	['US']	1.0	NaN	NaN
1	Taxi Driver	MOVIE	1976	R	113	['crime', 'drama']	['US']	NaN	tt0075314	8.3
2	Monty Python and the Holy Grail	MOVIE	1975	PG	91	['comedy', 'fantasy']	['GB']	NaN	tt0071853	8.2
3	Life of Brian	MOVIE	1979	R	94	['comedy']	['GB']	NaN	tt0079470	8.0
4	The Exorcist	MOVIE	1973	R	133	['horror']	['US']	NaN	tt0070047	8.1
...	...	...	...	...	...	...	...	...	...	...
5801	Fine Wine	MOVIE	2021	NaN	100	['romance', 'drama']	['NG']	NaN	tt13857480	6.9
5802	Edis Starlight	MOVIE	2021	NaN	74	['music', 'documentation']	[]	NaN	NaN	NaN
5803	Clash	MOVIE	2021	NaN	88	['family', 'drama']	['NG', 'CA']	NaN	tt14620732	6.5
5804	Shadow Parties	MOVIE	2021	NaN	116	['action', 'thriller']	[]	NaN	tt10168094	6.2
5805	Mighty Little Bheem: Kite Festival	SHOW	2021	NaN	0	['family', 'comedy', 'animation']	[]	1.0	tt13711094	8.8

Figure 1 : Tableau des données avant modification

Notre fichier Excel est formé par 11 colonnes et 5806 lignes, voici plus des détails :

Titre, Type (Spectacle ou film), Année de sortie (1953 jusqu'à 2022), Genres, votes …

Plusieurs colonnes qui ne sont pas nécessaires vont être supprimées comme l'id et la durée, de plus des autres modifications seront appliquées tels que supprimer les doublons, changer le format de certains caractères, diviser certaines colonnes et réinitialiser l'index.

La table finale est représentée dans l'image ci-dessous :

	title	type	release_year	imdb_score	imdb_votes	genres
0	Taxi Driver	MOVIE	1976	8.3	795222.0	crime
1	Monty Python and the Holy Grail	MOVIE	1975	8.2	530877.0	comedy
2	Life of Brian	MOVIE	1979	8.0	392419.0	comedy
3	The Exorcist	MOVIE	1973	8.1	391942.0	horror
4	Dirty Harry	MOVIE	1971	7.7	153463.0	thriller
...	...	...	...	...	...	...
3753	Momshies! Your Soul is Mine	MOVIE	2021	5.8	26.0	comedy
3754	Fine Wine	MOVIE	2021	6.9	39.0	romance
3755	Edis Starlight	MOVIE	2021	NaN	NaN	music
3756	Clash	MOVIE	2021	6.5	32.0	family
3757	Shadow Parties	MOVIE	2021	6.2	9.0	action

3758 rows × 6 columns

Figure 2 : Tableau des données après modifications

Chapitre II : Analyse exploratoire des données :

L'analyse exploratoire des données est une étape cruciale dans tout projet d'analyse de données. Il s'agit d'un processus itératif d'investigation et de visualisation de données pour découvrir des modèles, des relations et des anomalies potentielles.

Et dans ce chapitre, nous avons trouvé des informations intéressantes avec plus de détails dans le tableau ci-dessous concernant les votes:

Moyenne	**26683.21**
Médiane	2632
Médiane pondérée	2846
Premier quantile	537
Troisième quantile	12281

Et voici des autres intéressantes visualisations :

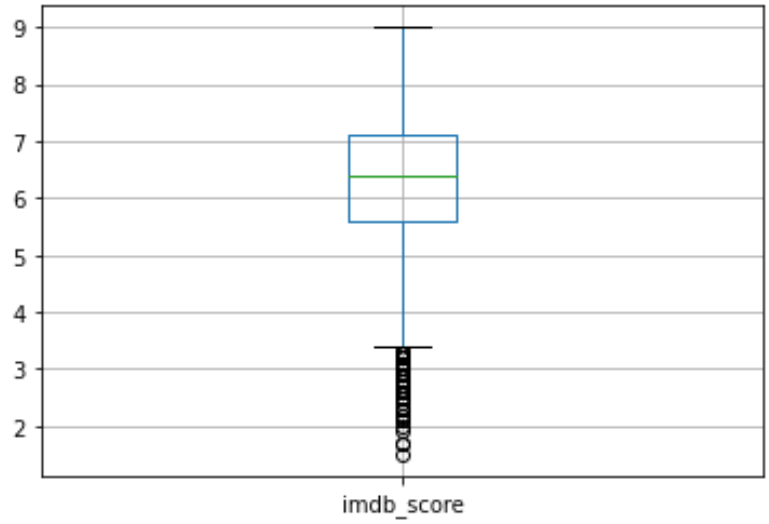

Figure 3 : Boîte à moustaches des scores

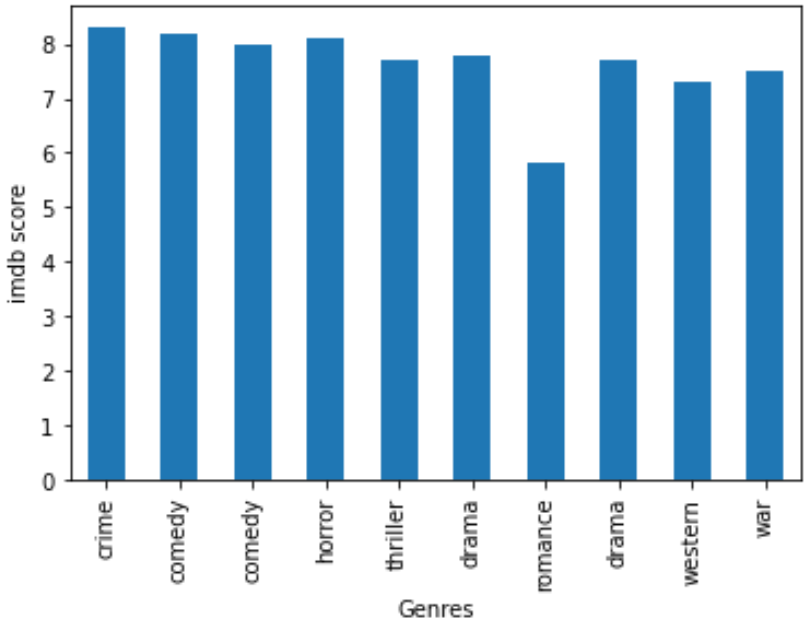

Figure 4 : Imdb score en fonction des Genres

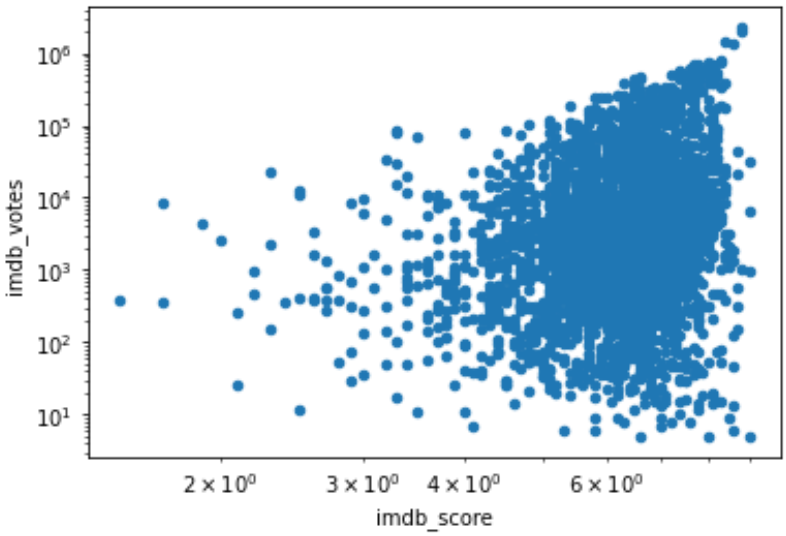

Figure 5 : Nuage de points, Score en fonction des votes

Chapitre III : Distribution des données et de l'échantillonnage :

En statistiques, une distribution d'échantillonnage est la distribution de probabilité d'une statistique (une statistique est un résumé de données, comme la moyenne ou l'écart type) calculée à partir de tous les échantillons possibles d'une certaine taille tirés d'une population. Cela montre essentiellement la probabilité que différentes valeurs de la statistique se produisent dans ces échantillons.

Voici des visualisations intéressantes :

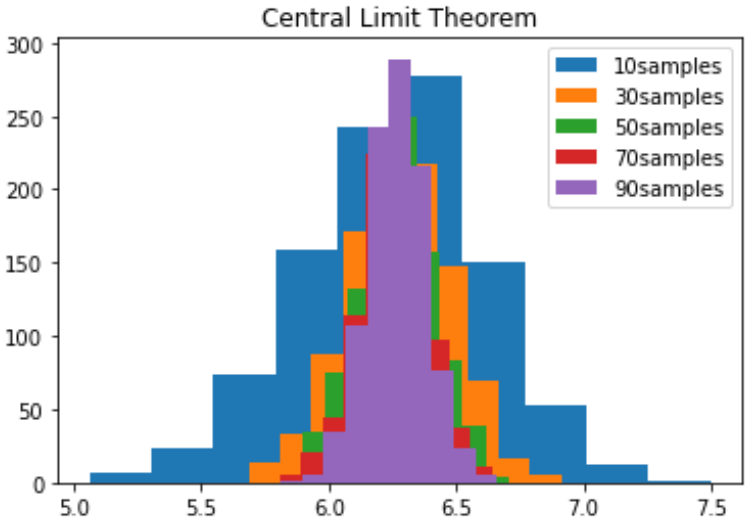

Figure 6 : Théorème central limite et échantillons

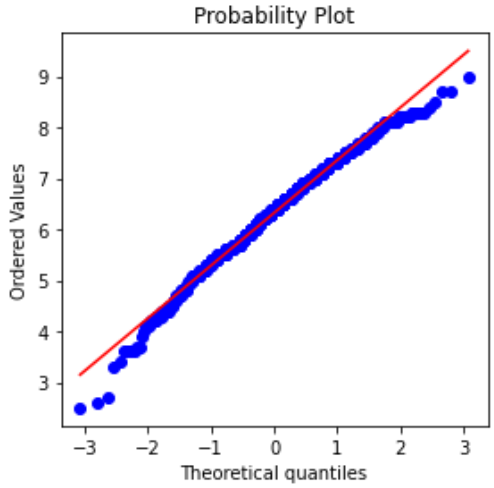

Figure 7 : Quantiles théoriques et valeurs ordonnées