Ministère de l'Enseignement Supérieur & de la Recherche Scientifique

Université de Carthage

FACULTE DES SCIENCES ECONOMIQUES ET DE GESTION DE NABEUL

MASTER PROFESSIONNEL EN INGÉNIERIE ECONOMIQUE ET FINANCIERE

Mini Projet :

Aide à la décision avec python

MATIERE : logiciel python

Elaborée par : Mezlini Rim

Année Universitaire 2023/2024

Plan

Introduction générale

L'industrie cinématographique est un écosystème dynamique, captivant et en constante évolution, où la diversité culturelle et la créativité convergent pour créer des expériences visuelles uniques. Ce rapport se plonge dans l'analyse approfondie d'un ensemble de données cinématographiques, explorant des aspects variés tels que la distribution du contenu, les tendances temporelles, et la contribution des différents pays à la production cinématographique.

À travers une combinaison de techniques d'analyse de données et de visualisation graphique, nous avons scruté les détails des films et séries TV présents dans le corpus. Des outils puissants tels que Python et ses bibliothèques dédiées ont été mobilisés pour extraire des insights significatifs, dévoilant des tendances, des préférences et des schémas culturels au sein de cette riche base de données.

Au cours de cette exploration, nous avons mis en lumière des éléments clés tels que les fluctuations temporelles dans l'ajout de contenu, les pays dominants dans la production cinématographique, et la préférence du public entre séries TV et films. Ces analyses servent de toile de fond pour une compréhension approfondie de l'industrie du divertissement et ouvrent la voie à des réflexions prospectives sur les évolutions à venir.

Plongeons maintenant dans les résultats détaillés de cette analyse, révélant les nuances et les histoires fascinantes inscrites dans les données cinématographiques que nous avons explorées.

Chapitre 1 : présentation des données

Présentation de l’entreprise :

Netflix est une entreprise américaine de divertissement fondée en 1997 par **Reed Hastings** et **Marc Randolph**. Initialement axée sur la location de DVD par la poste, elle a évolué pour devenir l'un des principaux services de streaming de contenu audiovisuel au monde. Netflix propose une vaste bibliothèque de films, séries télévisées, documentaires et programmes originaux dans différents genres.

Le modèle d'affaires de Netflix repose sur un abonnement mensuel, permettant aux utilisateurs de visionner du contenu en streaming illimité sur divers appareils, tels que des téléviseurs, des ordinateurs, des tablettes et des smartphones. L'entreprise a investi massivement dans la production de contenu original, créant des séries et des films exclusifs qui ont rencontré un grand succès critique et commercial.

Netflix est devenu un acteur majeur de l'industrie du divertissement, révolutionnant la façon dont les gens consomment des contenus audiovisuels et élevant le niveau de compétition dans le secteur du streaming. Avec une présence mondiale, Netflix continue d'innover et d'expérimenter de nouvelles formes de narration, devenant ainsi l'un des services de streaming les plus populaires et influents dans le monde entier.

L’intérêt :

Le choix des données de Netflix pour travailler sur un projet présente plusieurs avantages en raison de la nature riche et diversifiée de ses informations. Voici quelques-uns des principaux avantages :

Volume de données significatif : Netflix gère une énorme quantité de données liées à la consommation de médias, aux préférences des utilisateurs, aux évaluations, aux tendances de visionnage, etc. Travailler avec un ensemble de données de cette envergure offre des opportunités pour des analyses approfondies.

Diversité du contenu : La bibliothèque de Netflix englobe un large éventail de genres, de langues et de formats. Cela permet d'explorer des tendances de consommation de médias diverses et de créer des recommandations ou des analyses pertinentes pour un public varié.

Recommandations personnalisées : Netflix utilise des algorithmes avancés pour recommander du contenu aux utilisateurs en fonction de leurs préférences passées. Travailler avec ces données peut offrir des perspectives sur la manière dont les systèmes de recommandation fonctionnent et comment ils peuvent être améliorés.

Impact culturel : En raison de sa popularité mondiale, les données de Netflix peuvent refléter des tendances culturelles et des comportements de visionnage à l'échelle mondiale. Cela peut être particulièrement intéressant pour les projets qui cherchent à comprendre les dynamiques culturelles contemporaines.

Innovation dans la production de contenu : Netflix investit massivement dans la production de contenu original. Les données liées à la réception de ce contenu peuvent être explorées pour comprendre ce qui fonctionne bien et ce qui ne fonctionne pas dans l'industrie du divertissement, fournissant des informations utiles pour les projets liés à la création de contenu

Défis techniques : Travailler avec les données de Netflix peut également représenter un défi technique stimulant en raison de la complexité des algorithmes de recommandation, de la gestion du contenu à grande échelle, et d'autres aspects liés à la distribution de médias en streaming.

A propos du data set utilisé :

Les données collectées ont été transférées à un fichier Excel pour leurs chargement et exploration. Nous avons fait usage de la bibliothèque pandas, facilitant ainsi une exploration détaillée de leur organisation et de leur composition

```python
import pandas as pd
chemin_fichier_excel = 'C:\\netflix_titles_nov_2019.xlsx'
df = pd.read_excel(chemin_fichier_excel)
```

Ensuite pour afficher les données on utilise la commande suivante :

```python
df.head(10)
```

Et afficher ce tableau en prend les 5 premières lignes

	show_id	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	type
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	1 Season	International TV Shows, Korean TV Shows, Roman...	Brought together by meaningful meals in the pa...	TV Show
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 min	Documentaries, International Movies	From Sierra de las Minas to Esquipulas, explor...	Movie
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 min	Comedies, Dramas, International Movies	A goofy copywriter unwittingly convinces the l...	Movie
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 min	Dramas, Independent Movies, International Movies	Arranged to marry a rich man, young Ada is cru...	Movie
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman,	Canada, United Kingdom	NaN	2019	TV-Y	2 Seasons	Kids' TV	Lovable pug Chip starts kindergarten,	TV Show

Le data set est formé de **5939 lignes et 15 colonnes** et offre une vue formelle sur le contenu de la Platform Netflix représenté par exemple :

- Son Id
- Titre
- Directeur
- Acteurs principaux
- Date d’ajout
- Année de production
- Durée
- Description
- Type

Chapitre 2 : visualisation des données

1.les commandes de base :

J’ai effectué les opérations de base je vais donner quelque exemple :

```python
df.sort_values(by='release_year')
```

Trier le Data Frame par la colonne 'release_year'

*cette commande permet de supprimer les lignes qui contiennent des valeurs manquantes dons il élimine ligne de (NaN)

```python
df.dropna()
```

*Cette commande permet de citer les colonnes

```python
print(df.columns)
```

```
Index(['show_id', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description',
       'type', 'year_added', 'month_added', 'season_count'],
      dtype='object')
```

J’ai utilisé la bibliothèque NumPy et la bibliothèque pandas pour effectuer une opération de regroupement et de comptage sur un Data Frame basé sur la colonne 'country'

```python
import numpy as np
dd3 = df.groupby('country')
dd4 = dd3.count()
print(dd4)
```

```
                                                show_id  title  director  \
country
Argentina                                            38     38        26
Argentina, Brazil, France, Poland, Germany, Den...    1      1         1
Argentina, Chile                                      1      1         1
Argentina, Chile, Peru                                1      1         1
Argentina, France                                     1      1         1
...                                                 ...    ...       ...
Uruguay, Spain, Mexico                                1      1         1
Venezuela                                             1      1         1
Venezuela, Colombia                                   1      1         1
Vietnam                                               4      4         4
West Germany                                          1      1         1

                                                cast  date_added  \
country
Argentina                                         37          33
Argentina, Brazil, France, Poland, Germany, Den...  1           1
Argentina, Chile                                    1           1
Argentina, Chile, Peru                              1           1
Argentina, France                                   1           1
...                                               ...         ...
Uruguay, Spain, Mexico                              1           1
Venezuela                                           1           1
Venezuela, Colombia                                 0           1
...
Vietnam                                                   4     4
West Germany                                              1     1

[528 rows x 11 columns]
```

2.Visualisation graphique :

La visualisation graphique est cruciale dans l'analyse de données, offrant une puissante manière de représenter des informations complexes de manière visuelle. Cette section explore les graphiques générés avec Python, une plateforme polyvalente et largement utilisée en science des données. Ces représentations vont au-delà de simples illustrations, elles sont des outils essentiels pour révéler des tendances, des schémas et des insights au sein des données analysées. Des bibliothèques telles que **Matplotlib, Seaborn** et **Plotly** offrent une diversité d'outils graphiques permettant une personnalisation approfondie, enrichissant ainsi l'exploration et l'interprétation des données. Plongeons dans cet univers captivant des graphiques Python pour comprendre comment ces représentations deviennent des alliés indispensables dans l'analyse approfondie des données.

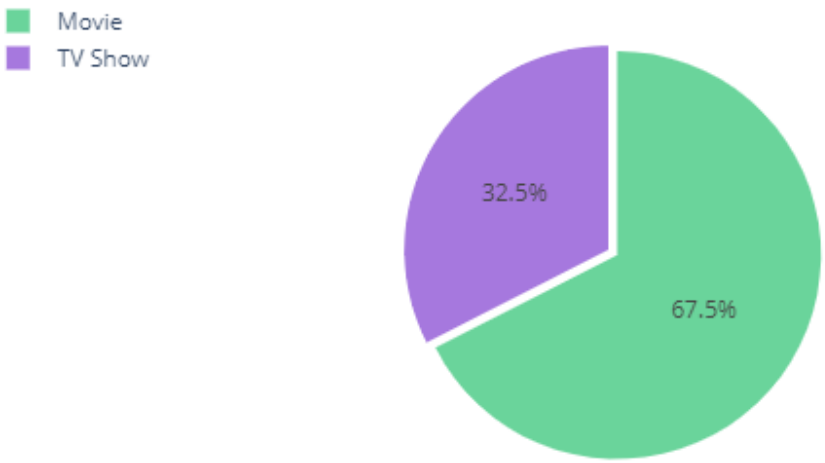
Je vais donner quelque exemple des graphiques que j’ai présenté :

```python
import plotly.graph_objects as go
from plotly.offline import iplot, plot

col = "type"
grouped = df[col].value_counts().reset_index()
grouped = grouped.rename(columns={col: "count", "index": col})


trace = go.Pie(labels=grouped[col], values=grouped['count'], pull=[0.05, 0], marker=dict(colors=["#6ad49b", "#a678de"]))
layout = go.Layout(title="", height=400, legend=dict(x=0.1, y=1.1))
fig = go.Figure(data=[trace], layout=layout)

iplot(fig)
```

Python

Cette série de commandes en Python utilise la bibliothèque Plotly pour créer un diagramme circulaire (pie chart) basé sur les données du Data Frame

---



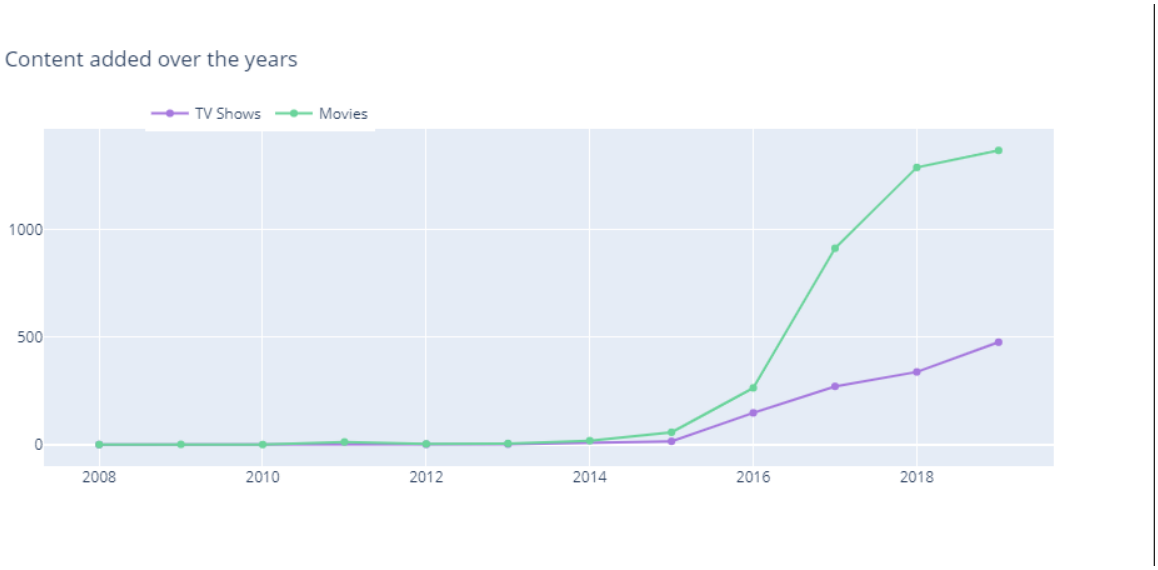
Ce graphique met en évidence une répartition significative entre les séries TV et les films, fournissant une vue rapide et informative de la composition du contenu dans l'ensemble de données.

Ensuite j’ai utilisé la bibliothèque Plotly pour créer un graphique à dispersion (scatter plot) montrant l'évolution du contenu ajouté au fil des années, en distinguant entre les séries TV et les films dans le Data Frame

```python
d1 = df[df["type"] == "TV Show"]
d2 = df[df["type"] == "Movie"]

col = "year_added"

vc1 = d1[col].value_counts().reset_index()
vc1 = vc1.rename(columns = {col : "count", "index" : col})
vc1['percent'] = vc1['count'].apply(lambda x : 100*x/sum(vc1['count']))
vc1 = vc1.sort_values(col)

vc2 = d2[col].value_counts().reset_index()
vc2 = vc2.rename(columns = {col : "count", "index" : col})
vc2['percent'] = vc2['count'].apply(lambda x : 100*x/sum(vc2['count']))
vc2 = vc2.sort_values(col)

trace1 = go.Scatter(x=vc1[col], y=vc1["count"], name="TV Shows", marker=dict(color="#a678de"))
trace2 = go.Scatter(x=vc2[col], y=vc2["count"], name="Movies", marker=dict(color="#6ad49b"))
data = [trace1, trace2]
layout = go.Layout(title="Content added over the years", legend=dict(x=0.1, y=1.1, orientation="h"))
fig = go.Figure(data, layout=layout)
fig.show()
```

Python

Donc il nous donne ce graphique :



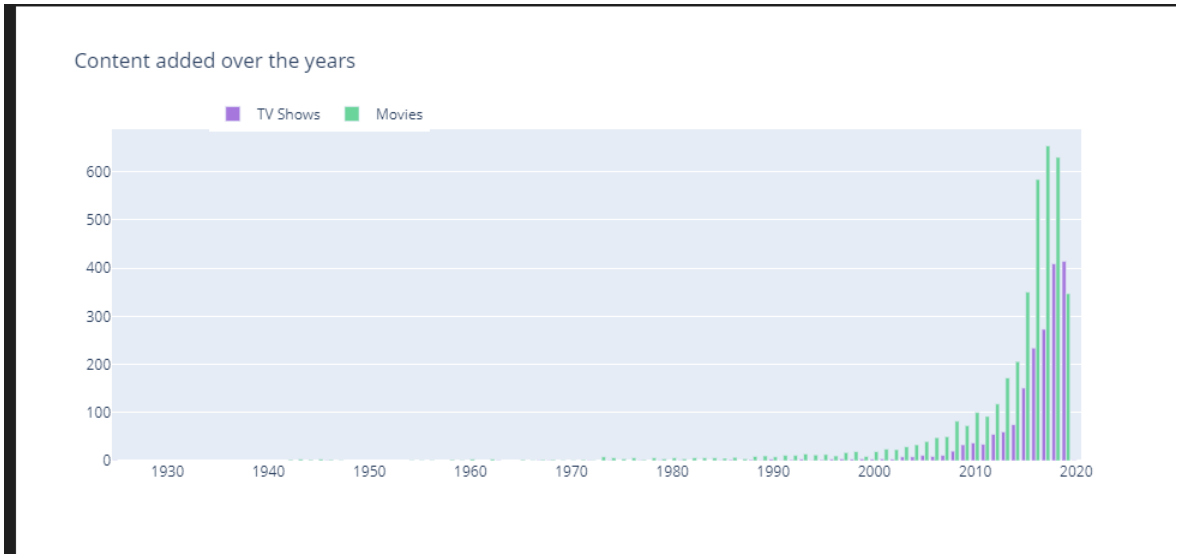
Ce graphique offre un aperçu visuel dynamique de l'évolution temporelle des films et les séries du 2008 à 2018, permettant des observations sur les tendances générales, les différences entre les séries TV et les films, ainsi que des points saillants liés aux pics et aux fluctuations d'ajouts de contenu au fil des années.

Ensuite j’ai utilisé la bibliothèque Plotly pour créer un graphique à barres (bar chart) illustrant l'évolution du contenu ajouté au fil des années pour les séries TV et les films dans les Data Frames d1 et d2

```python
col = "release_year"

vc1 = d1[col].value_counts().reset_index()
vc1 = vc1.rename(columns = {col : "count", "index" : col})
vc1['percent'] = vc1['count'].apply(lambda x : 100*x/sum(vc1['count']))
vc1 = vc1.sort_values(col)

vc2 = d2[col].value_counts().reset_index()
vc2 = vc2.rename(columns = {col : "count", "index" : col})
vc2['percent'] = vc2['count'].apply(lambda x : 100*x/sum(vc2['count']))
vc2 = vc2.sort_values(col)

trace1 = go.Bar(x=vc1[col], y=vc1["count"], name="TV Shows", marker=dict(color="#a678de"))
trace2 = go.Bar(x=vc2[col], y=vc2["count"], name="Movies", marker=dict(color="#6ad49b"))
data = [trace1, trace2]
layout = go.Layout(title="Content added over the years", legend=dict(x=0.1, y=1.1, orientation="h"))
fig = go.Figure(data, layout=layout)
fig.show()
```

Python

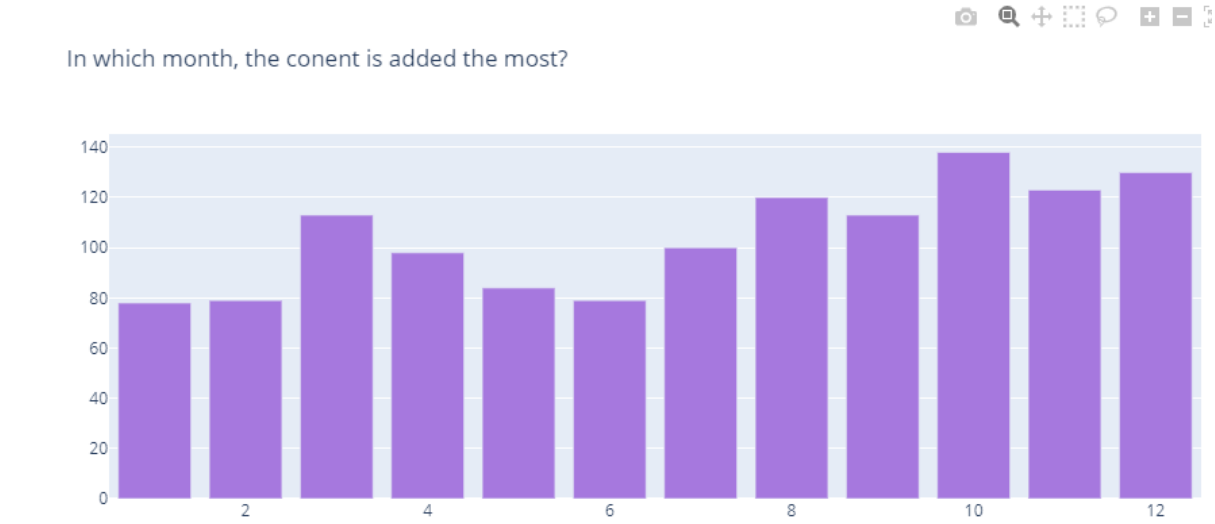
Ce graphique à barres offre une représentation visuelle des tendances temporelles du contenu ajouté, permettant une comparaison entre les séries TV et les films au fil des années. Donc on remarque la tendance des films depuis 1970 or que les séries des TV apparaisse en 2000 sur un échelle de 1930 à 2020



Puis j’ai fait au cours de quel mois les films et les série des TV sont le plus ajoutés alors j’ai utilisé la bibliothèque Plotly pour créer un graphique à barres (bar chart) illustrant la répartition du contenu ajouté pour les séries TV au fil des mois.

```python
col = 'month_added'
vc1 = d1[col].value_counts().reset_index()
vc1 = vc1.rename(columns = {col : "count", "index" : col})
vc1['percent'] = vc1['count'].apply(lambda x : 100*x/sum(vc1['count']))
vc1 = vc1.sort_values(col)

trace1 = go.Bar(x=vc1[col], y=vc1["count"], name="TV Shows", marker=dict(color="#a678de"))
data = [trace1]
layout = go.Layout(title="In which month, the conent is added the most?", legend=dict(x=0.1, y=1.1, orientation="h"))
fig = go.Figure(data, layout=layout)
fig.show()
```

Python



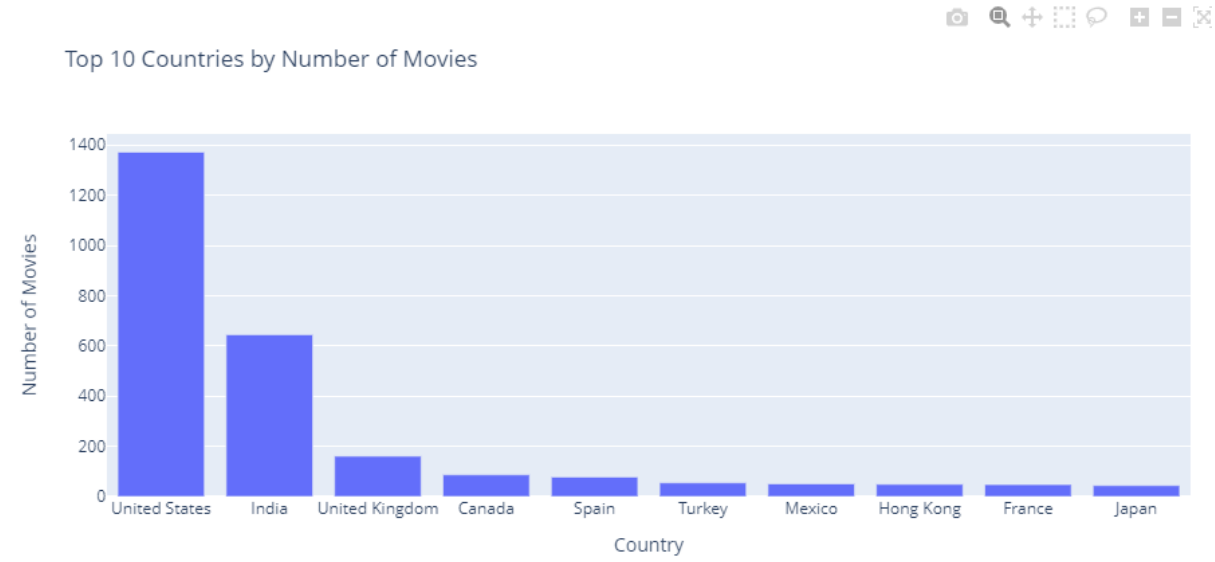
J’ai travaillée sur Top 10 des pays par nombre de films

```python
movies_df = df[df['type'] == 'Movie']
country_counts = movies_df['country'].value_counts().reset_index()
country_counts.columns = ['Country', 'Number of Movies']
top_10_countries = country_counts.head(10)
fig = go.Figure([go.Bar(x=top_10_countries['Country'], y=top_10_countries['Number of Movies'])])
fig.update_layout(title='Top 10 Countries by Number of Movies',
                  xaxis_title='Country',
                  yaxis_title='Number of Movies')
fig.show()
```

Python

Ce script utilise la bibliothèque Plotly pour créer un graphique à barres (bar chart) illustrant le nombre de films par pays

---



Ce graphique à barres présente une vue claire des 10 pays ayant produit le plus grand nombre de films dans le Data Frame original, donc on remarque les deux premier qui produits les plus est l’United states et India offrant ainsi une indication visuelle des principaux contributeurs à la production cinématographique

Conclusion

En clôture de cette analyse approfondie des données cinématographiques, plusieurs constats émergent, offrant une vision éclairante de l'industrie du divertissement. L'exploration des tendances temporelles a révélé une dynamique en constante évolution, avec des années marquantes et des périodes de prolifération dans la production cinématographique et télévisuelle.

La distinction entre séries TV et films a été scrutée, mettant en lumière une préférence prononcée pour l'un ou l'autre. La visualisation graphique a permis de dresser le portrait des principaux contributeurs à la production mondiale de films, révélant des concentrations géographiques intéressantes.

Au-delà des chiffres, cette analyse a capturé la richesse culturelle présente dans les données, reflétant la diversité des contenus cinématographiques à travers le monde. Les pays émergents comme des acteurs clés dans cette narration globale, ajoutant des nuances et des perspectives uniques à l'industrie cinématographique mondiale.

En conclusion, cette exploration a dévoilé la puissance des données dans la compréhension de l'écosystème cinématographique. Les insights obtenus ne sont pas simplement des chiffres, mais des fragments d'histoires, des reflets de sociétés, et des indicateurs des préférences culturelles. Ce rapport ouvre la voie à des discussions continues sur les évolutions futures de cette industrie, stimulant la curiosité pour les prochaines intrigues que les données cinématographiques pourraient dévoile