Netfix insights

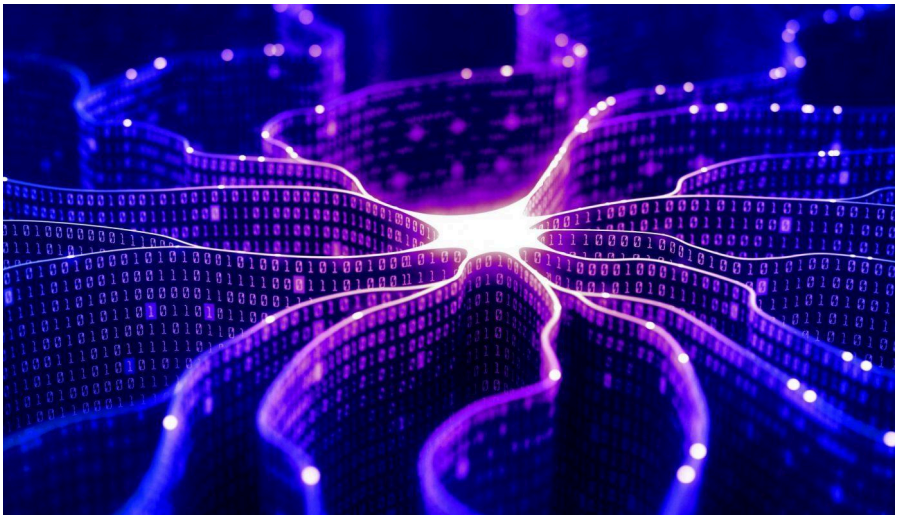
“Chris Rock's new comedy special just came out on Netflix. It slaps.”

AI Watch

Voulant pousser votre apprentissage de l’intelligence artificielle, vous réalisez **une veille** sur les différents aspects suivants. Définissez simplement ce qu’est :

- **L’intelligence artificielle,**
- **Le Machine Learning (ou l’apprentissage automatique),**
- **Le pré-traitement des données,**
- **L’analyse descriptive des données.**

Choisissez **un domaine** (la santé, la finance, l’environnement, le juridique, l’immobilier, etc) et **expliquez ce qu’il est possible de faire** dans ce domaine grâce à l’intelligence artificielle.

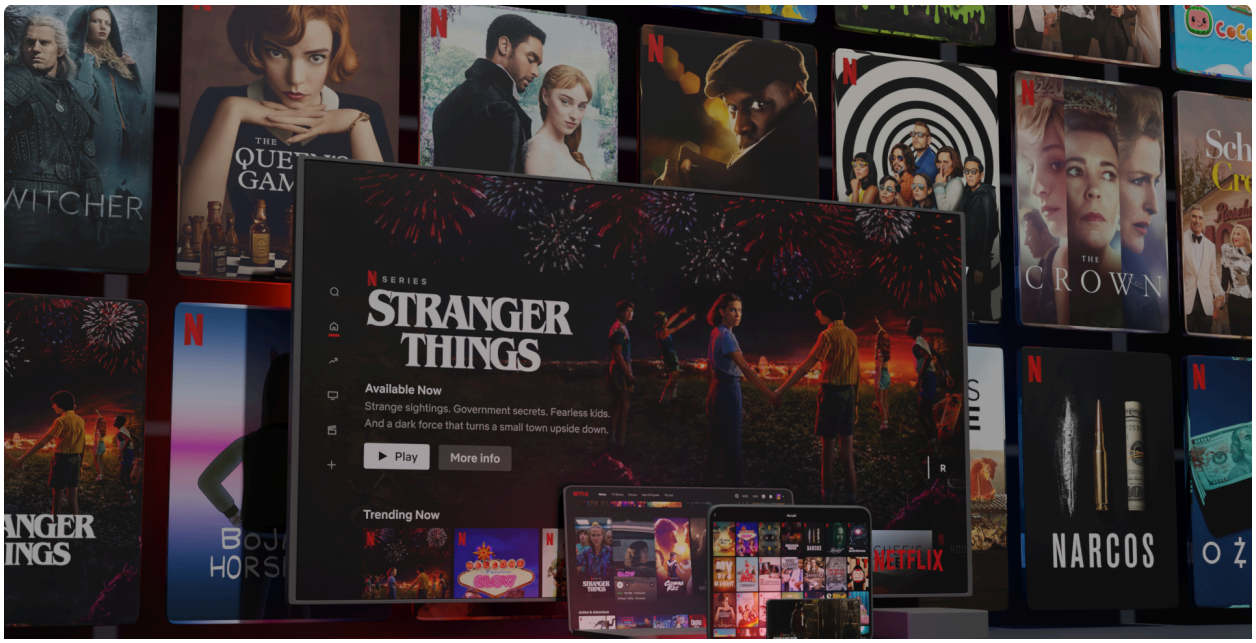


World No. 2 in Streaming & Online TV

Netflix, géant mondial du divertissement, a transformé la façon dont nous consommons les films et les émissions de télévision. Depuis sa création en 1997 en tant que service d'envoi de DVD par courrier, Netflix n'a cessé d'évoluer, s'adaptant aux avancées technologiques et devenant finalement la force dominante de l'industrie du streaming. Avec plus de 200 millions d'abonnés dans le monde, Netflix offre une vaste bibliothèque de contenus et propose des recommandations personnalisées, captivant les spectateurs avec une expérience de streaming transparente.

Nourrissant un intérêt grandissant pour l'intelligence artificielle, vous décidez de vous lancer dans une étape importante, l'analyse de données ! Vous posez les premières fondations de votre apprentissage en répondant aux questions ci-dessous.

Vous analyserez les données des séries TV et des films sur Netflix datant de septembre 2021, à récupérer ici.



1. Réalisez une veille sur l'outil **Jupyter Notebook** et installez le sur votre machine (soit en passant par Anaconda, soit directement sur VSCode). Familiarisez vous avec les cellules de code ainsi que les cellules de Markdown.

2. **Créez** et ouvrez **un notebook Jupyter**. Nommez le "Netflix Data Analysis".

3. Chargez le dataset à l'aide de **Pandas** sous le format d'un DataFrame.

4. **Affichez un aperçu du DataFrame** en affichant ses 5 premières observations et ses 5 dernières observations à l'aide de deux fonctions spécifiques de Pandas.

5. **Affichez les informations du DataFrame**, notamment le type d'index et les colonnes, les valeurs non nulles et l'utilisation de la mémoire.

6. Affichez **la dimensionnalité du DataFrame**. Combien avez-vous de variables ? Combien avez-vous d'observations ?

7. Affichez **les colonnes du DataFrame**.

8. Affichez le **type des différentes colonnes du DataFrame**. Avez-vous des données quantitatives (numériques) ? Avez-vous des données qualitatives (catégorielles) ? Si oui, lesquelles ?

9. **Y a-t-il des données manquantes** ? Identifiez la proportion en pourcentage.

10. A l'aide de la librairie **Missingno**. Affichez **un graphique représentatif de la proportion** des données manquantes.

11. Affichez **une observation aléatoire** du DataFrame.

12. Affichez **toutes les informations** de l'œuvre **"Catch Me If You Can"**.

13. Affichez le nom du **film le plus récent** du dataset.

14. Affichez le nom de **la série la plus récente** du dataset.

15. Modifiez la variable **date_added** de telle sorte qu'elle soit de **type DateTime**.

16. Modifiez la variable **duration** de telle sorte que la durée des films soit un nombre plutôt qu'une chaîne de caractères, par exemple : **160 au lieu de "160 min"**.

17. Modifiez la variable **duration** de telle sorte que la durée des séries soit un nombre plutôt qu'une chaîne de caractères, par exemple : **2 au milieu de "2 seasons"**.

18. Modifiez la variable **listed_in** de telle sorte que la chaîne de caractères soit un liste de chaîne de caractères, par exemple : **["International TV Shows", "TV Dramas", "TV Mysteries"] au lieu de "International TV Shows, TV Dramas, TV Mysteries"**.

19. Affichez les **valeurs uniques** des variables : **type, country, release_year, rating et listed_in**.

20. Voyez-vous un "**director**" ayant produit **plus d'une œuvre** ?
21. Quelle est l'année avec **le plus de films ajoutés au catalogue de Netflix** ?
22. Quelle est l'année avec **le plus de séries ajoutées au catalogue de Netflix** ?

23. **Visualisez vos données à l'aide des différentes librairies de Matplotlib, Seaborn ou Plotly** :

    a. La répartition du type d'oeuvres du dataset,

    b. La répartition des oeuvres en fonction des pays du dataset,

    c. La répartition des années du dataset,

    d. La répartition des ratings du dataset,

    e. La répartition de la durée des films du dataset,

f. La répartition de la durée des séries du dataset,

g. La répartition des genres d'œuvres du dataset,

h. Le top 5 des séries les plus longues,

i. Le top 5 des films les plus longs,

j. La répartition des “directors” des œuvres françaises,

k. La répartition des œuvres en fonction de la date d’ajout au catalogue.

l. Tout autre graphe **pertinent** pour vous.

24. Prenez soin d’**observer et analyser** tous vos graphiques. **Que pouvez-vous dire sur vos données** ?

Compétences visées

➔ Analyse de données

Rendu

Dans un repository github public nommé **netflix-insights**, vous devrez fournir les éléments suivants :

1. **Un Notebook Jupyter** avec votre analyse des données de Netflix. N’oubliez pas d’ajouter des titres, une introduction, un sommaire et une conclusion.
2. **Un README.md** associé à votre repository, contenant :
    a. **La veille réalisée** de la partie AI Watch.

b. Une présentation du **contexte du projet** ainsi que les données utilisées.

c. Les **différentes observations et les conclusions** de votre analyse des données de Netflix.

Base de connaissances

- Pandas Documentation
- Python | Visualize missing values (NaN)
- Matplotlib: Visualization with Python
- Seaborn: statistical data visualization
- Plotly Open Source Graphing Library for Python
- A Guide To Getting Data Visualization Right
- Data Visualization Reference Guides