

Analyse de Données Netflix - Documentation

Mohamed Rashad Rouine

November 11, 2023

1 Introduction

Le projet d'analyse de données Netflix vise à explorer et comprendre le jeu de données en utilisant Python et Streamlit. Cette initiative a pour objectif de fournir des analyses approfondies et des visualisations informatives des films originaux de Netflix. L'outil développé permet d'interagir avec les données, de filtrer les résultats par genre et langue, et de présenter des statistiques clés ainsi que des graphiques illustrant l'évolution des scores IMDb, la distribution des durées, et plus encore.

2 Aperçu du Code

Le cœur du projet consiste à charger un jeu de données Netflix, à effectuer une analyse de données et à créer des visualisations avec Streamlit. Le code Python peut être consulté ci-dessous :

```
1 \begin{lstlisting}[caption={Explication du Code}, label=code:explication]
2 import pandas as pd
3 import streamlit as st
4 import matplotlib.pyplot as plt
5
6 # Chargement du jeu de données Netflix depuis un fichier CSV
7 file_path = 'NetflixOriginals.csv'
8
9 # Tentative de chargement avec l'encodage 'utf-8'
10 try:
11     df = pd.read_csv(file_path, encoding='utf-8')
12 except UnicodeDecodeError:
13     # Si 'utf-8' choue, essayer avec 'latin-1' ou 'ISO-8859-1'
14     df = pd.read_csv(file_path, encoding='latin-1')
15
16 # Conversion de la colonne 'Premiere' au format datetime
17 df['Premiere'] = pd.to_datetime(df['Premiere'], errors='coerce', format='%d
    -%b-%y')
18
19 # Mise en page de l'application Streamlit avec un titre
20 st.title("Tableau de bord des films Netflix")
21
22 # Affichage des données brutes sous le sous-titre 'Données brutes'
23 st.subheader('Données brutes')
24 st.write(df)
25
26 # Élément interactif : Filtre par genre et langue
27 selected_genre = st.selectbox('Sélectionnez un genre:', ['Tous'] + list(df
    ['Genre'].unique()))
28 selected_language = st.selectbox('Sélectionnez une langue:', ['Toutes'] +
    list(df['Language'].unique()))
29
30 # Filtrage des données en fonction du genre et de la langue sélectionnés
31 filtered_data = df.copy()
32 if selected_genre != 'Tous':
```

```

33     filtered_data = filtered_data[filtered_data['Genre'] == selected_genre]
34 if selected_language != 'Toutes':
35     filtered_data = filtered_data[filtered_data['Language'] ==
        selected_language]
36
37 # Graphique 1 : Affichage des statistiques pour le genre et la langue
    s lectionn s
38 st.subheader(f"Statistiques pour le genre {selected_genre} et la langue {
        selected_language} :")
39 st.write(filtered_data[['IMDB Score', 'Runtime']].describe())
40
41 # Graphique 2 : Affichage du Top 5 des films par score IMDB
42 st.subheader('Top 5 des films par score IMDB')
43 if not filtered_data.empty:
44     top_5_movies = filtered_data.nlargest(5, 'IMDB Score')[['Title', 'IMDB
        Score']]
45     st.table(top_5_movies.set_index('Title'))
46
47 # Graphique 3 : Affichage du graphique d'volution pour le score IMDB
48 st.subheader('L\'volution pour le score IMDB')
49 if not filtered_data.empty:
50     imdb_evolution = filtered_data.groupby('Premiere')['IMDB Score'].mean()
51     st.line_chart(imdb_evolution)
52
53 # Graphique 4 : Affichage de l'histogramme pour la dur e des films
54 st.subheader('La distribution des dur es')
55 fig, ax = plt.subplots()
56 ax.hist(filtered_data['Runtime'], bins=20, edgecolor='black')
57 ax.set_xlabel('Dur e des films')
58 ax.set_ylabel('Fr quence des films')
59 st.pyplot(fig)
60
61 # Graphique 5 : Affichage de l'Area Chart pour la distribution des films par
    date de sortie au fil du temps (cumulatif)
62 st.subheader('La distribution des films par date de sortie au fil du temps (
        cumulatif)')
63 if not filtered_data.empty:
64     release_date_distribution = filtered_data.groupby('Premiere').size().
        cumsum().reset_index(name='Cumulative Count')
65     st.area_chart(release_date_distribution.set_index('Premiere'))

```

Listing 1: Code d'Analyse de Données Netflix

3 Problèmes Rencontrés

Malgré les progrès réalisés, j'ai rencontré plusieurs défis lors du développement du projet. Voici un résumé des principaux problèmes et de leurs résolutions :

3.1 Problème d'Encodage Unicode

Le jeu de données a présenté une erreur `UnicodeDecodeError` en raison de problèmes d'encodage du texte. Pour résoudre cela, j'ai utilisé différents encodages tels que `'utf-8'`, `'latin-1'`, et `'ISO-8859-1'`.

3.2 Problème de Conversion de Date

Le format de date de la colonne `'Premiere'` n'était pas reconnu initialement. J'ai résolu ce problème en spécifiant le format correct lors de la conversion de la colonne avec `pd.to_datetime()`.

3.3 Problèmes des Graphiques

J'ai rencontré des difficultés avec le chevauchement des étiquettes du graphique en secteurs (pie chart) et le fonctionnement incorrect de l'histogramme. Le problème du graphique en secteurs a été résolu en affichant uniquement les deux premières lettres de chaque langue et en ajustant la taille de la police. L'histogramme est toujours en cours d'amélioration.

3.4 Problème de Déploiement

En tentant de déployer l'application, j'ai rencontré une erreur indiquant que le fichier n'était pas reconnu comme un fichier `.py`. J'ai souhaité intégrer l'application sur l'un de mes sites web, similaire à mes projets précédents sur TBS (<https://www.easystem.tn/rashad/>). J'ai poussé le code sur GitHub (accessible à [https://github.com/MohamedRashadRouine/python/tree/main/BI%20\(python\)/DS%20%20project](https://github.com/MohamedRashadRouine/python/tree/main/BI%20(python)/DS%20%20project)). Malheureusement, je continue de faire face à la même erreur, laissant penser que le fichier n'est pas reconnu en tant que fichier `.py`. Pour plus de détails, vous pouvez consulter le journal des erreurs dans le fichier `logs`.

4 Conclusion

Le projet d'Analyse de Données Netflix a été une expérience d'apprentissage significative, et avec l'aide de ChatGPT, j'ai amélioré mes compétences en codage Python. Alors que je continue à affiner le projet et à surmonter les défis, j'ai hâte d'améliorer ma maîtrise de l'analyse et de la visualisation de données.

Remarque : J'ai principalement compté sur ChatGPT tout au long de ce projet, étant encore en apprentissage du langage Python. Chaque ligne de code m'a permis d'acquérir de nouvelles compétences.