Compte Rendu de Projet Entrepôt de Données

Date : 12 Avril 2023

Nom du Projet : Analyse de Marché pour l'Investissement de Mr Smith

Membres de l'équipe : Miri Rayen , Kouki Ayhem , Elee Ben Ameur

# 1. Introduction

Ce projet vise à développer un système informatique décisionnel pour aider notre client, Mr Smith, à analyser le marché des jeux vidéo, des livres et des films. L'objectif est de fournir un dashboard interactif qui guidera Mr Smith dans sa décision d'investissement.

# 2. Avancement du Projet

Recherche de Données :

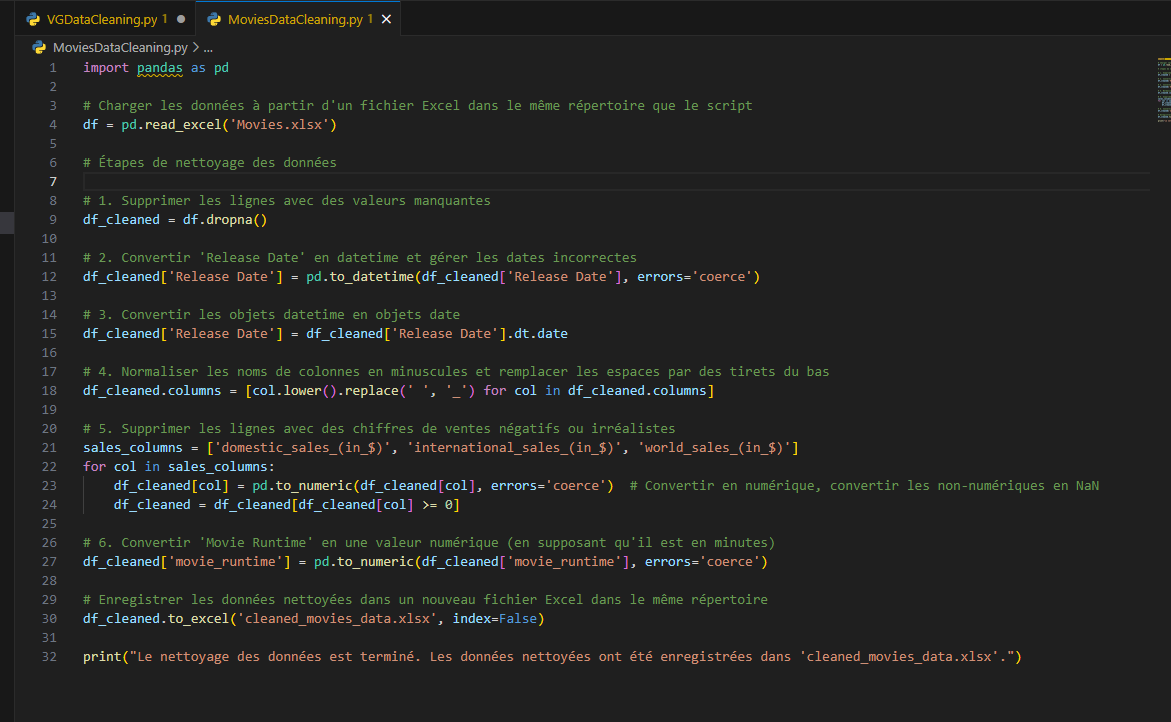
Nous avons commencé par examiner les trois jeux de données fournis. Pour éviter les problèmes liés aux fichiers plats, nous avons opté pour des fichiers Excel, qui nous ont permis de mieux gérer les données séparées par des virgules.

Data Flow et ETL :

Le data flow a été établi en utilisant SSIS, avec une attention particulière à la transformation des données Excel pour garantir une intégration fluide dans notre base de données SQL Server. Bien que nous ayons rencontré des défis initiaux avec le mappage des types de données, ces problèmes ont été résolus par des ajustements dans le processus d'importation.

Data Cleaning avec Python :

Nous avons écrit un script Python pour nettoyer les données, en supprimant les valeurs nulles, en corrigeant les erreurs de formatage et en éliminant les caractères indésirables. Voici un extrait du code utilisé pour ces opérations(MoviesDataCleaning) :



Explication du code :  
Après le chargement des données, nous effectuons plusieurs étapes de nettoyage des données. Tout d'abord, nous supprimons les lignes contenant des valeurs manquantes en utilisant la fonction dropna().

Ensuite, nous convertissons la colonne 'Release Date' en format datetime et gérons les dates incorrectes en utilisant la fonction pd.to\_datetime() avec l'argument 'coerce'. Cela convertira toutes les dates invalides en valeurs NaN.

Nous convertissons ensuite les objets datetime en objets date en utilisant l'attribut dt.date.

Ensuite, nous normalisons les noms de colonnes en minuscules et remplaçons les espaces par des tirets du bas en utilisant une liste de compréhension.

Après cela, nous supprimons les lignes contenant des chiffres de ventes négatifs ou irréalistes. Nous définissons d'abord une liste de colonnes contenant des chiffres de ventes, puis itérons sur chaque colonne et la convertissons en format numérique en utilisant la fonction pd.to\_numeric() avec l'argument 'coerce'. Cela convertira toutes les valeurs non numériques en NaN. Nous filtrons ensuite les lignes où les chiffres de ventes sont inférieurs à 0.

Enfin, nous convertissons la colonne 'Movie Runtime' en une valeur numérique (en supposant qu'elle est en minutes) en utilisant la fonction pd.to\_numeric() avec l'argument 'coerce'.

Conception de l'Entrepôt de Données :

Nous avons décidé de classer les données en dimensions pour les colonnes constantes et en faits pour les colonnes variables. Cela a été fait en prévision de la création de cubes OLAP pour l'analyse multidimensionnelle.

# 3. Résultats et Observations

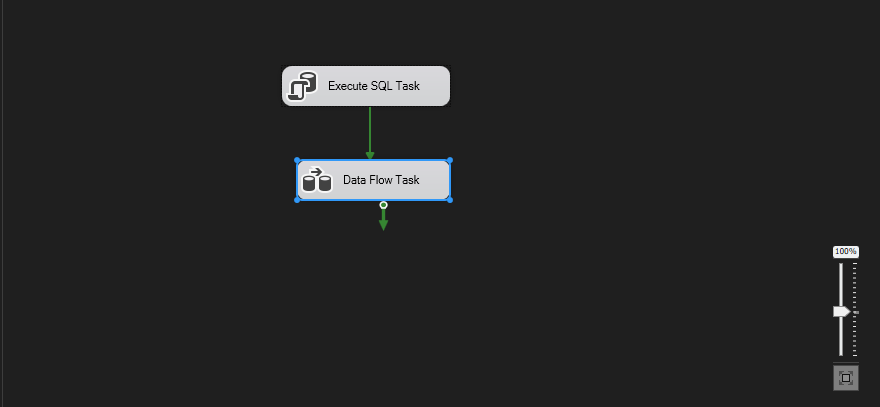
Les données ont été nettoyées avec succès et sont prêtes pour l'étape suivante de la modélisation. Le processus ETL est en place et fonctionne comme prévu. Nous avons une bonne compréhension de la structure de notre entrepôt de données, ce qui facilitera la création du dashboard.

# 4. Conclusion

Nous sommes sur la bonne voie pour répondre aux besoins de Mr Smith. Les étapes de nettoyage et de préparation des données ont été réalisées avec succès, et la conception de l'entrepôt de données est bien avancée. Nous sommes confiants dans notre capacité à livrer un dashboard informatif et décisionnel dans les délais impartis.

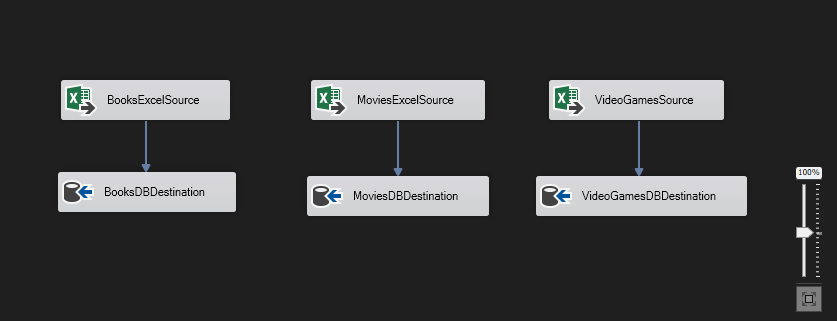
# 5.Annexe

Capture d’écran du Control Flow



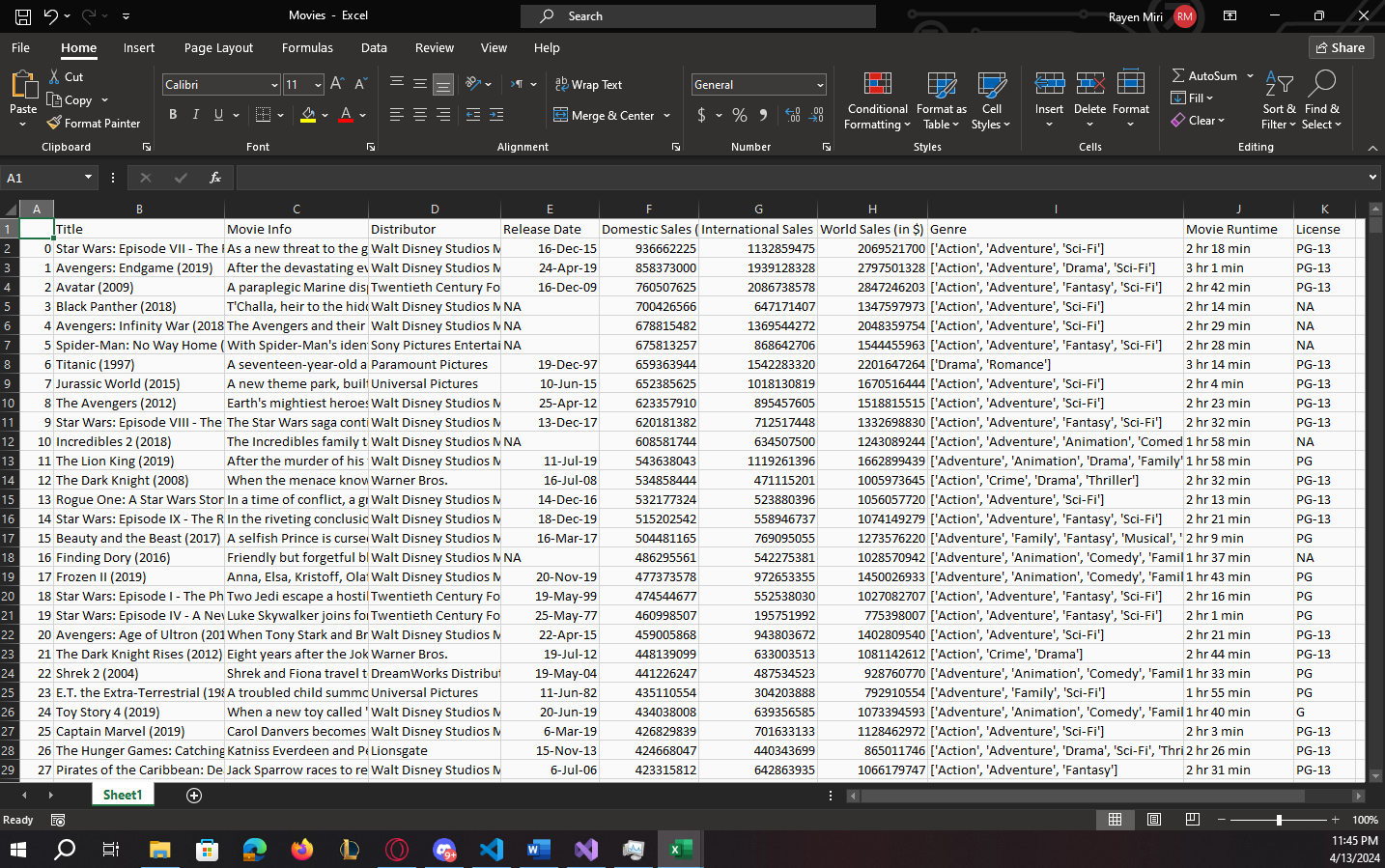
Control Flow 1

Capture d’écran du Data Flow



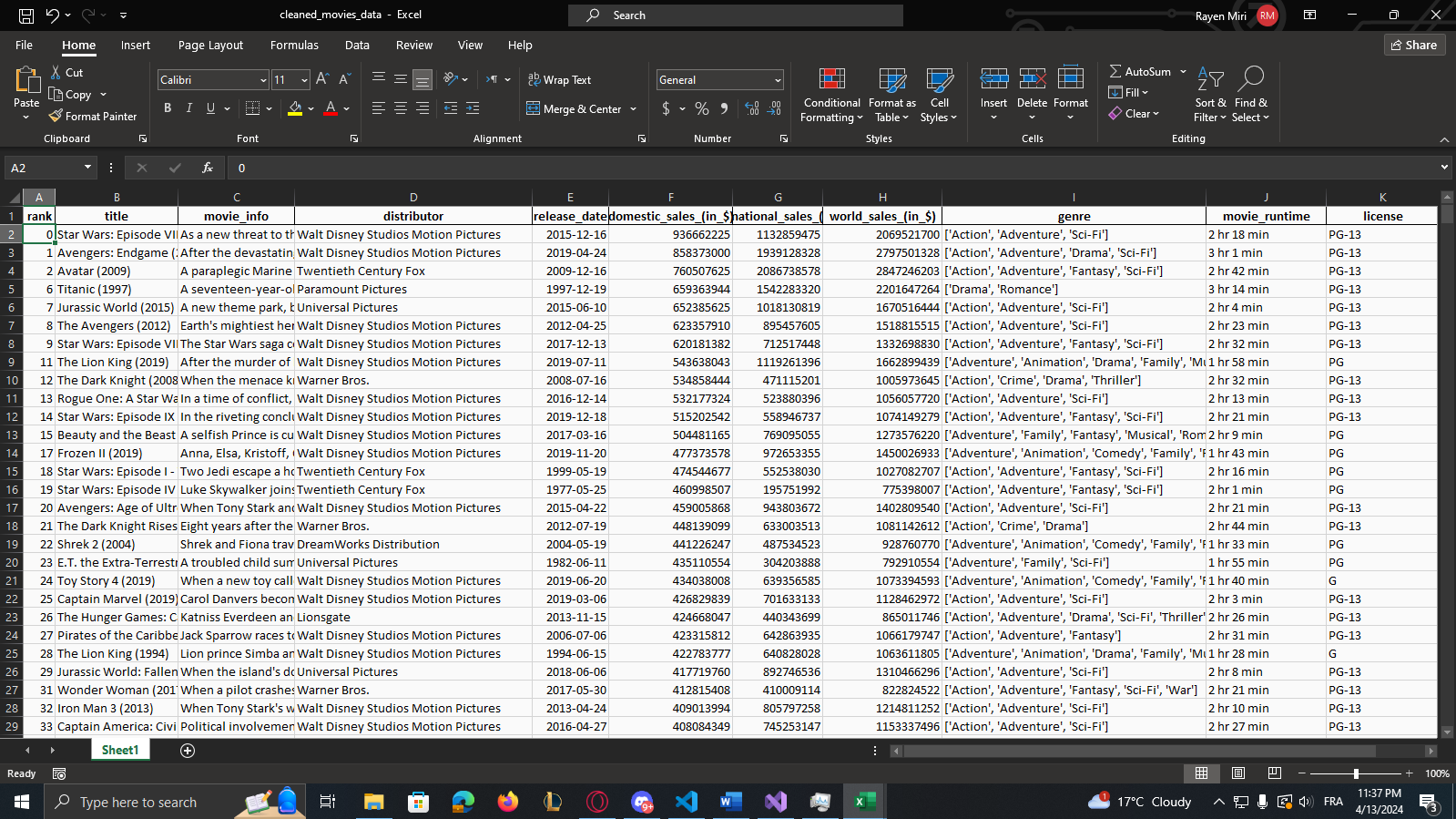
Data Flow 1

Capture d’écran du Data Avant et Après le DataCleaning



Movies Data Avant de Data Cleaning 1

MoviesData Avant de Data Cleaning 2



Movies Data Appès le Data Cleaning 1

# 7.Liste des figures

[Control Flow 1 3](file:///C:\Users\GLADIATEUR-1650-8\Desktop\ProjetEntropot\Compte%20Rendu%20de%20Projet%20%20%20%20Entrepôt%20de%20Données.docx#_Toc163944677)

[Data Flow 1 4](file:///C:\Users\GLADIATEUR-1650-8\Desktop\ProjetEntropot\Compte%20Rendu%20de%20Projet%20%20%20%20Entrepôt%20de%20Données.docx#_Toc163944678)

[Movies Data Avant de Data Cleaning 1 4](file:///C:\Users\GLADIATEUR-1650-8\Desktop\ProjetEntropot\Compte%20Rendu%20de%20Projet%20%20%20%20Entrepôt%20de%20Données.docx#_Toc163944679)

[MoviesData Avant de Data Cleaning 2 4](file:///C:\Users\GLADIATEUR-1650-8\Desktop\ProjetEntropot\Compte%20Rendu%20de%20Projet%20%20%20%20Entrepôt%20de%20Données.docx#_Toc163944680)

[Movies Data Appès le Data Cleaning 1 5](file:///C:\Users\GLADIATEUR-1650-8\Desktop\ProjetEntropot\Compte%20Rendu%20de%20Projet%20%20%20%20Entrepôt%20de%20Données.docx#_Toc163944681)