import pandas as pd

import re

import ast

from sklearn.preprocessing import MultiLabelBinarizer

from sklearn.preprocessing import LabelEncoder

#Duplicates

df=pd.read\_csv('games.csv')

df.columns = df.columns.str.lower().str.replace(' ', '\_')

def convert\_k\_to\_number(x):

    if isinstance(x, str) and 'K' in x:

        return float(x.replace('K', '')) \* 1000

    return x

for col in ['times\_listed', 'number\_of\_reviews', 'plays', 'playing', 'backlogs', 'wishlist']:

    if col in df.columns:

        df[col] = df[col].apply(convert\_k\_to\_number)

df['release\_date'] = pd.to\_datetime(df['release\_date'], errors='coerce')

df['days\_since\_release'] = (df['release\_date'] - df['release\_date'].min()).dt.days

df['release\_date']= df['days\_since\_release']

df.duplicated().sum()

df.drop\_duplicates(inplace=True)

df.to\_csv('games.csv',index=False)

#split to numerical and categorical

df\_cat=df.select\_dtypes(include='object')

df\_num=df.select\_dtypes(exclude='object')

print(df\_cat.columns)

print(df\_num.columns)

#find null values

df\_num.isnull().sum()

df\_cat.isnull().sum()

#find outliers

def outlier\_capping(x):

  q1=x.quantile(0.25)

  q3=x.quantile(0.75)

  iqr=q3-q1

  lower=q1-1.5\*iqr

  upper=q3+1.5\*iqr

  x=x.clip(lower=lower,upper=upper)

  return x

for i in df\_num.columns:

  df\_num[i]=outlier\_capping(df\_num[i])

for i in ["title","summary","reviews"] :

    encoder=LabelEncoder()

    df\_cat[i]=encoder.fit\_transform(df\_cat[i])

def parse\_genres(genre\_string):

    try:

        return ast.literal\_eval(genre\_string)

    except (ValueError, SyntaxError):

        return []

print(df\_cat['genres'])

df\_cat['genres'] = df\_cat['genres'].apply(parse\_genres)

mlb = MultiLabelBinarizer()

genres\_encoded = pd.DataFrame(mlb.fit\_transform(df\_cat['genres']), columns=mlb.classes\_, index=df.index)

df\_cat = df\_cat.join(genres\_encoded)

df\_cat = df\_cat.drop('genres', axis=1)

merged\_df = df\_num.join(df\_cat)

merged\_df.to\_csv('cleaned\_games.csv', index=False)

print("Merged DataFrame created and saved to 'cleaned\_games.csv'")