import pandas as pd

import warnings

warnings.filterwarnings("ignore")

#do not change the predefined function names

#Task 1: Remove columns that are not needed in our analysis.

# Remove Url\_spotify, Uri, Key, Url\_youtube, Description

def Remove\_columns():

#do not remove following line of code

df = pd.read\_csv('Spotify\_Youtuben.csv')

#WRITE YOUR CODE HERE

df.drop(columns=['Url\_spotify','Uri','Key','Url\_youtube','Description'], inplace=True)

#return dataframe

return df

#Task 2: Check for the null values

def no\_of\_null\_values():

#Do not remove the following code statment

df=Remove\_columns()

#WRITE YOUR CODE HERE TO CHECK THE NO OF NULL VALUES AND RETURNS THE SAME

df=df.isnull().sum()

#return sum of null values by columns

return df

#Task 3: Handle the null values replace int value with 0 and other values with NA

def Handle\_Null\_values():

#Do not remove the following code statment

df=Remove\_columns()

#WRITE YOUR CODE HERE ACCORDING TO THE DESCRIPTION

#df=df.fillna(value={col: 0 if df[col].dtype==int else 'NA' for col in df.columns})

df=df.fillna({'Artist':'NA', 'Track':'NA','Album':'NA','Album\_type':'NA', 'Danceability':0,'Energy':0,'Loundness':0,'Speechiness':0,'Acounticness':0,'Instrumentalness':0,'Liveness':0,'Valence':0,'Tempo':0,'Duration\_ms':0,'Title':'NA','Channel':'NA','Views':0,'Likes':0,'Comments':0,'Licensed':'NA','offical\_video':'NA','Stream':0})

#return dataframe

return df

#Task 4: CHECK FOR DUPLICATES AND REMOVE THEM KEEPING THE FIRST VALUE

def drop\_the\_duplicates():

#Do not remove the following code statment

df=Handle\_Null\_values()

#WRITE YOUR CODE HERE

df=df.drop\_duplicates() # (keep='first')

#return dataframe

return df

#Task 5: CONVERT millisecond duration to minute for a better understanding

def convert\_milisecond\_to\_Minute():

#Do not remove the following code statment

df=drop\_the\_duplicates()

#WRITE YOUR CODE HERE

# df['Duration\_ms']=pd.to\_numeric(df['Duration\_ms'],errors='coerce')

df['Duration\_ms']=df['Duration\_ms']/60000

return df

#Task 6: Rename the modified column to Duration\_min

def rename\_modified\_column():

#Do not remove the following code statment

df=convert\_milisecond\_to\_Minute()

#WRITE YOUR CODE HERE

df=df.rename(columns={'Duration\_ms':'Duration\_min'})

#return dataframe

return df

#Task 7: Remove irrelevant 'Track' name that starts with ?

def Irrelevant\_Track\_name():

#Do not remove the following code statment

df=rename\_modified\_column()

#WRITE YOUR CODE HERE

#df['Track']=df['Track'].apply(lambda x:x[1:] if x.startswith('?')else x)

#df=df[df['Track'].str.strip() !='']

df=df[~df['Track'].str.startswith('?')]

#return dataframe

return df

#Task 8: Calculate the Energy to Liveness ratio for each track and store it in columns 'EnergyLiveness'

def Energy\_to\_liveness\_Ratio():

#Do not remove the following code statment

df=Irrelevant\_Track\_name()

#WRITE YOUR CODE HERE

#df['Energy']=pd.to\_numeric(df['Energy'],errors='coerce')

#df['Liveness']=pd.to\_numeric(df['Liveness'],errors='coerce')

df['EnergyLiveness']=df['Energy'] / df['Liveness']

#return dataframe

return df

#Task 9: change the datatype of 'views' to float for further use

def change\_the\_datatype():

#Do not remove the following code statment

df=Energy\_to\_liveness\_Ratio()

#WRITE YOUR CODE HERE

#df['Views']=df['Views'].replace('NA',float('nan'))

df['Views']=df['Views'].astype(float)

#return dataframe

return df

#Task 10: compare the views and stream columns to infer

# that the song track was more played on which platform, youtube or Spotify.

# Create a column named most\_playedon which will have two values.

# Spotify and Youtube,If a song track is most played on youtube then

# the most\_played on column will have youtube as the value for that particular song

def compare\_the\_views():

#Do not remove the following code statment

df=change\_the\_datatype()

#WRITE YOUR CODE HERE

#df['Views']=pd.to\_numeric(df['Views'],errors='coerce')

#df['Stream']=pd.to\_numeric(df['Stream'],errors='coerce')

#df['diff']=abs(df['Stream']-df['Views'])

#df['Ratio']=df['Views']/df['Stream']

#threshold=0.67

df['most\_playedon']=df.apply(lambda x: 'Spotify' if x['Stream'] >x['Views'] else 'Youtube',axis=1)

#return dataframe

return df

#Task 11: export the cleaned dataset to CSV to "cleaned\_dataset.csv"

def export\_the\_cleaned\_dataset():

#Do not remove the following code statment

df=compare\_the\_views()

#WRITE YOUR CODE HERE

dataset=df.to\_csv('cleaned\_dataset.csv')

#create csv file "cleaned\_dataset.csv" using dataframe

#TASK 12

#follow the instruction in the Task 13 description and complete the task as per it.

#check if mysql table is created using "cleaned\_dataset.csv"

#Use this final dataset and upload it on the provided database for performing analysis in MySQL

#To run this task click on the terminal and click on the run projec