import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

#open song-of-10s.csv and store its data in a DataFrame

df\_songs = pd.read\_csv("songs-of-10s.csv")

print(df\_songs.head())#print

print(df\_songs.shape)#rows by columns

print("total number of missing values:",df\_songs.isnull().sum())

df\_genres = pd.read\_csv("genres-of-10s.csv")

print(df\_genres.head())#print

print(df\_genres.shape)#rows by columns

df\_songs\_header =pd.read\_csv("songs-of-10s.csv",encoding='utf-8',header=1)

#DataFrame is Data framework(数据框架)

#df\_songs = pd.DataFrame(df\_songs,columns=df\_songs\_header)

df\_genres\_header = pd.read\_csv("genres-of-10s.csv",encoding='utf-8',header=1)

#df\_genres = pd.DataFrame(df\_genres,columns=df\_genres.header)

df\_songs.set\_index("artist",inplace=True)

df\_genres.set\_index("artist",inplace=True)

df\_sum= df\_songs.merge(df\_genres,on=["artist"],how="outer")

print(df\_sum)

#How many instances are there?

print("How many instances are there?")

df\_songs\_sum=df\_songs.shape;

print(df\_songs\_sum)

df\_genres\_sum=df\_genres.shape;

print(df\_genres\_sum)

#How many attributes are there?

#Which ones are categorical?

#Which ones are numeric

#Are there any missing values? If so how many and in which column(s)?

print("How many attributes are there? , Which ones are categorical? and Which ones are numeric ,Are there any missing values? If so how many and in which column(s)?")

df\_songs\_header.info()

df\_genres\_header.info()

print("which column(s) is null")

df\_songs\_isnull=df\_songs.isnull();

print(df\_songs\_isnull)

df\_genres\_isnull=df\_genres.isnull();

print(df\_genres\_isnull)

print("--------------------------------")

print(list(df\_songs))

print(list(df\_genres))

#How many hits are there and how many flops are there?

print(len(df\_songs["track"].value\_counts()))

print(df\_songs["track"].unique())

print(len(df\_genres["genres"].value\_counts()))

print(df\_genres["genres"].unique())

#What are the top 5 artists with the most songs in the dataset?

print("-----------sort---------")

df\_songs\_sort = df\_songs.sort\_values("uri", ascending=False).groupby("uri").head(5)

df\_genres\_sort = df\_genres.sort\_values("genres",ascending=False).groupby("genres").head(10)

print(df\_songs\_sort)

print(df\_genres\_sort)

#Part 2 Cleaning and Exploratory Data Analysis (40 pts)

#Clean/Prepare the Data

df\_songs.dropna(inplace=True)

print("after dropping rows with missing values")

print(df\_songs.isnull().sum())

print(df\_songs.shape)

#replace "1" with "hit" and "0" with "flop"

de\_songs\_clean=df\_songs.replace("hit",1)

de\_songs\_clean=df\_songs.replace("flop",0)

#out to a new file called dataset-of-10s-cleaned.csv

de\_songs\_clean.to\_csv("dataset-of-10s-cleaned.csv", index=False)

#Group the DataFrame by hit/flop

print("Group the DataFrame by hit/flop")

df\_songs\_cleaned = pd.read\_csv("dataset-of-10s-cleaned.csv")

songs\_by\_target = df\_songs\_cleaned.groupby("target")

songs\_target\_ser = songs\_by\_target["target"].sum()

genres\_by\_target = df\_genres\_sort.groupby("genres")

genres\_target\_ser = genres\_by\_target["genres"].sum()

print(songs\_target\_ser)