import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import LabelEncoder,OneHotEncoder

movies=pd.read\_csv("moviesL.csv")

def repl(str1):

return str1.replace('|',' ')

movies.genres=movies.genres.apply(repl)

movierate=[np.nan]\*len(movies)

movies=movies.assign(rating=movierate)

movies.set\_index('movieId',inplace=True)

movies

ratings=pd.read\_csv("ratingsL.csv")

del ratings['timestamp']

avg\_ratings=ratings.groupby('movieId').mean()

avg\_ratings.reset\_index(inplace = True)

del avg\_ratings['userId']

avg\_ratings.set\_index('movieId',inplace=True)

avg\_ratings

for i in avg\_ratings.index:

movies.rating[i]=avg\_ratings.rating[i]

movies.dropna(how='any',inplace=True)

genre\_list=['Action','Adventure','Animation','Children','Comedy','Crime','Documentary','Drama','Fantasy','Film-Noir','Horror','Musical','Mystery','Romance','Sci-Fi','Thriller','War','Western']

ls2=[]

for x in genre\_list:

ls1=[]

for i in movies.index:

if x in movies.genres[i]:

ls1.append(1)

else:

ls1.append(0)

ls2.append(ls1)

for i in range(0,len(genre\_list)):

movies[genre\_list[i]]=ls2[i]

from sklearn.preprocessing import StandardScaler

scaler=StandardScaler()

scaled\_movies=scaler.fit\_transform(movies[['rating', 'Action', 'Adventure',

'Animation', 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',

'Fantasy', 'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance',

'Sci-Fi', 'Thriller', 'War', 'Western']])

new\_movies=pd.DataFrame(scaled\_movies,columns=['rating', 'Action', 'Adventure',

'Animation', 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',

'Fantasy', 'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance',

'Sci-Fi', 'Thriller', 'War', 'Western'])

inertias=[]

for k in range(1,101):

print(k)

kmeans=KMeans(n\_clusters=k,init='k-means++')

kmeans.fit\_transform(new\_movies[['rating', 'Action', 'Adventure',

'Animation', 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',

'Fantasy', 'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance',

'Sci-Fi', 'Thriller', 'War', 'Western']])

inertias.append(kmeans.inertia\_)

plt.scatter(range(1,101),inertias)

plt.show()

kmeans=KMeans(n\_clusters=29,init='k-means++')

kmeans.fit\_transform(new\_movies[['rating', 'Action', 'Adventure',

'Animation', 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',

'Fantasy', 'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance',

'Sci-Fi', 'Thriller', 'War', 'Western']])

labels=list(kmeans.labels\_)

cluster\_size=[0]\*29

for i in range(0,len(labels)):

cluster\_size[labels[i]]+=1