Homework1

**现有从网络爬取的4916部电影基本信息，请用所学的Pandas数据分析处理方法完成以下任务**

**(1) 读取movie.csv文件并生成相应的Dataframe**

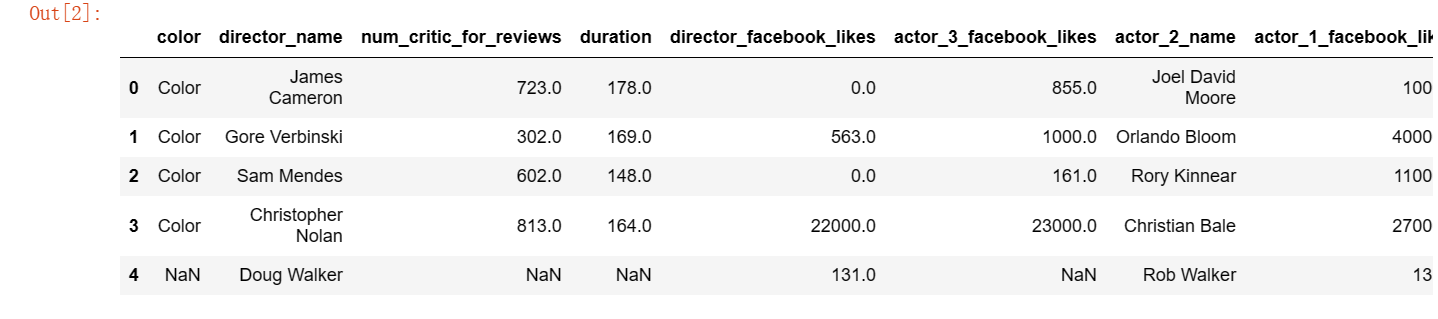
import pandas as pd

import numpy as np

df = pd.read\_csv("movie.csv")

df.head()

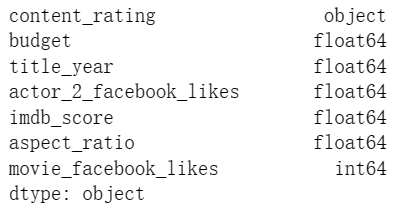
结果如下：



**（2）分析并输出该Dataframe中各列的数据类型**

df.dtypes

结果如下：



**(3) 应用数据Selection和loc (或iloc)方法选择所有类型为非数值的列，**

**并生成新的Dataframe**

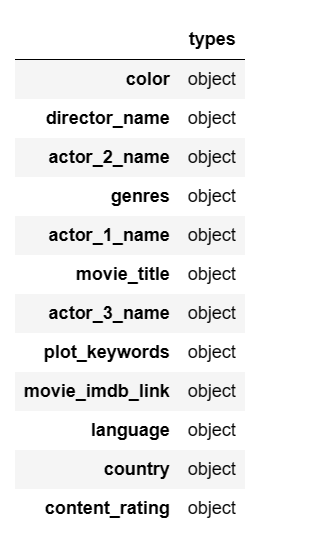
df1=df.dtypes

df1=pd.DataFrame(df1)

df1.columns=['types']

df1[(df1['types']=='object')]

结果如下：



**(4) 编写适当的数据filtering条件，将2000年以后美国出品的语言为**

**English的题材(genres)包含Action的彩色电影选择出来，并生成新的**

**Dataframe**

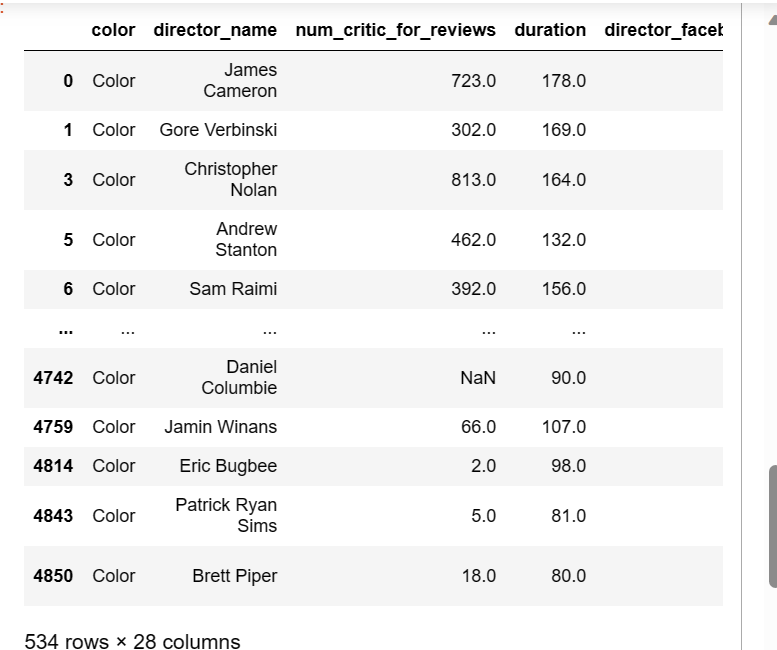
df[(df['title\_year']>2000)

&(df['country']=='USA')

&(df['language']=='English')

&(df['genres'].str.contains('Action'))]

结果如下：



**(5) 自定义函数并应用apply和applymap方法将缺失的数值填充为所在**

**列的平均值(mean)，将缺失的名义值填充为众数(mode)**

df1.loc[df1['types'] == 'object'] = df1.loc[df1['types'] == 'object'].applymap(lambda x: np.nan)

**(6) 应用GroupBy和agg方法按照电影的不同年份和题材分组，并对投**

**票的用户数(num\_voted\_users)和评论的用户数(num\_user\_for\_reviews)**

**进行汇总，并生成新的Datafra选择条件是内容分级为PG-13 ，并导出**

**为csv文件**

df2=df[(df.content\_rating=='PG-13')]

grouped\_by\_year = df\_pg\_13.groupby('title\_year').agg({

'num\_voted\_users': 'sum',

'num\_user\_for\_reviews': 'sum'

}).reset\_index()

grouped\_by\_genre = df\_pg\_13.groupby('genres').agg({

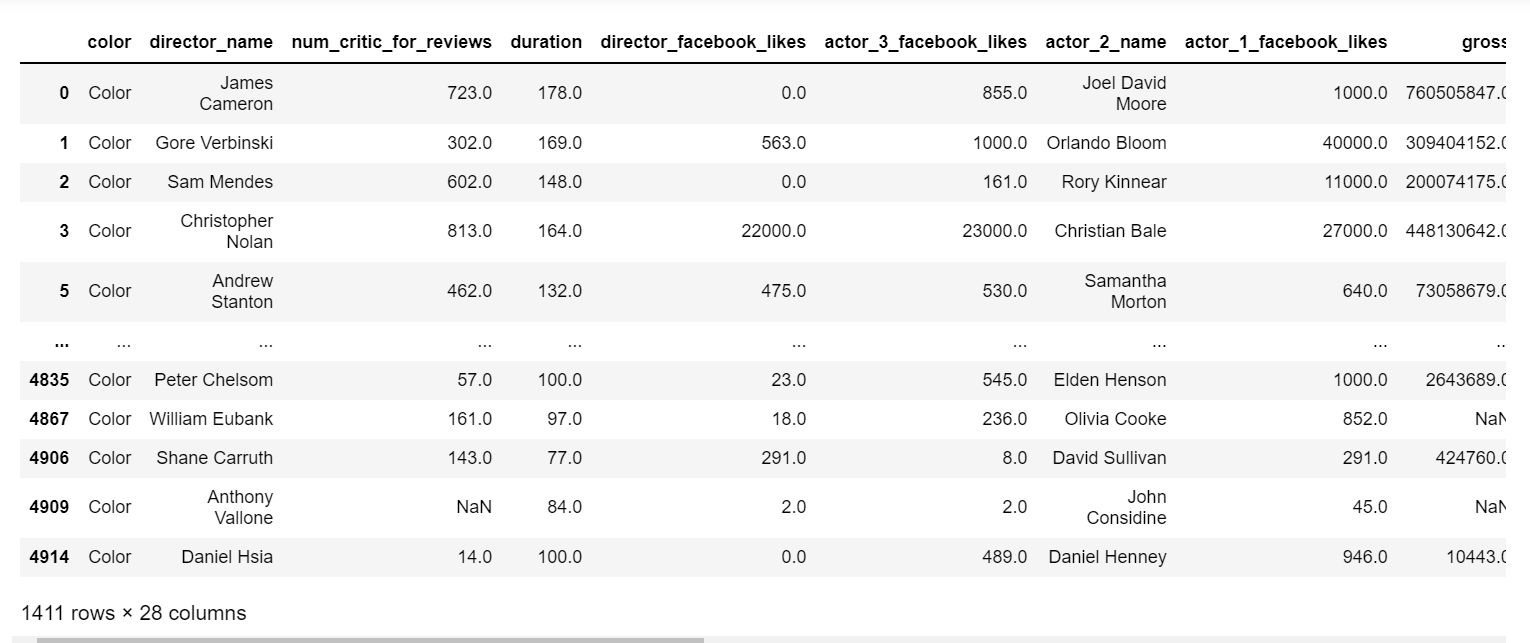
'num\_voted\_users': 'sum',

'num\_user\_for\_reviews': 'sum'

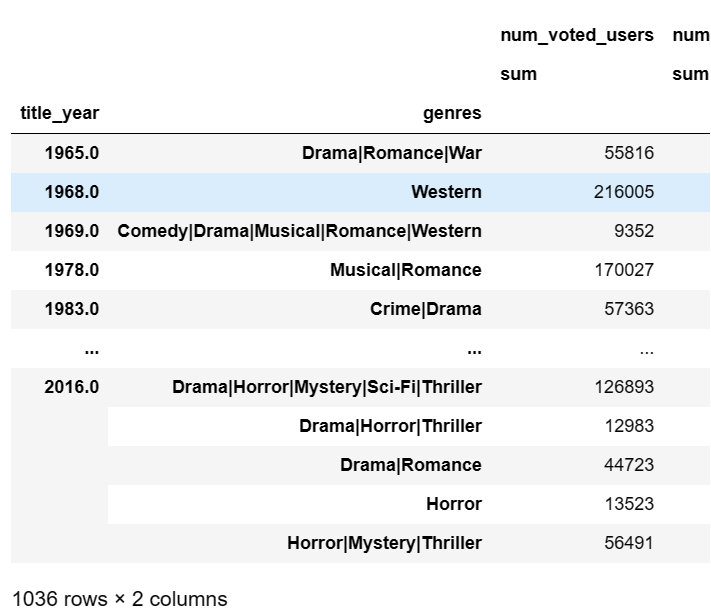
}).reset\_index()

df2.groupby(['title\_year','genres']).agg({'num\_voted\_users':['sum'],'num\_user\_for\_reviews':['sum']})

结果如下：







**(7) 应用Pandas可视化方法生成不同国家出品的内容分级为PG-13的电**

**影数量的柱状图，并将其保存为png文件。（提示：选择条件是内容**

**分级为PG-13，横轴是国家，纵轴是数量）**

import matplotlib.pyplot as plt

df3=df[(df.content\_rating=='PG-13')]

df3.groupby(['country']).agg({'country':['count']})

number=df3.groupby(['country']).agg({'country':['count']})

number.plot(kind='bar')

plt.show

结果如下：



