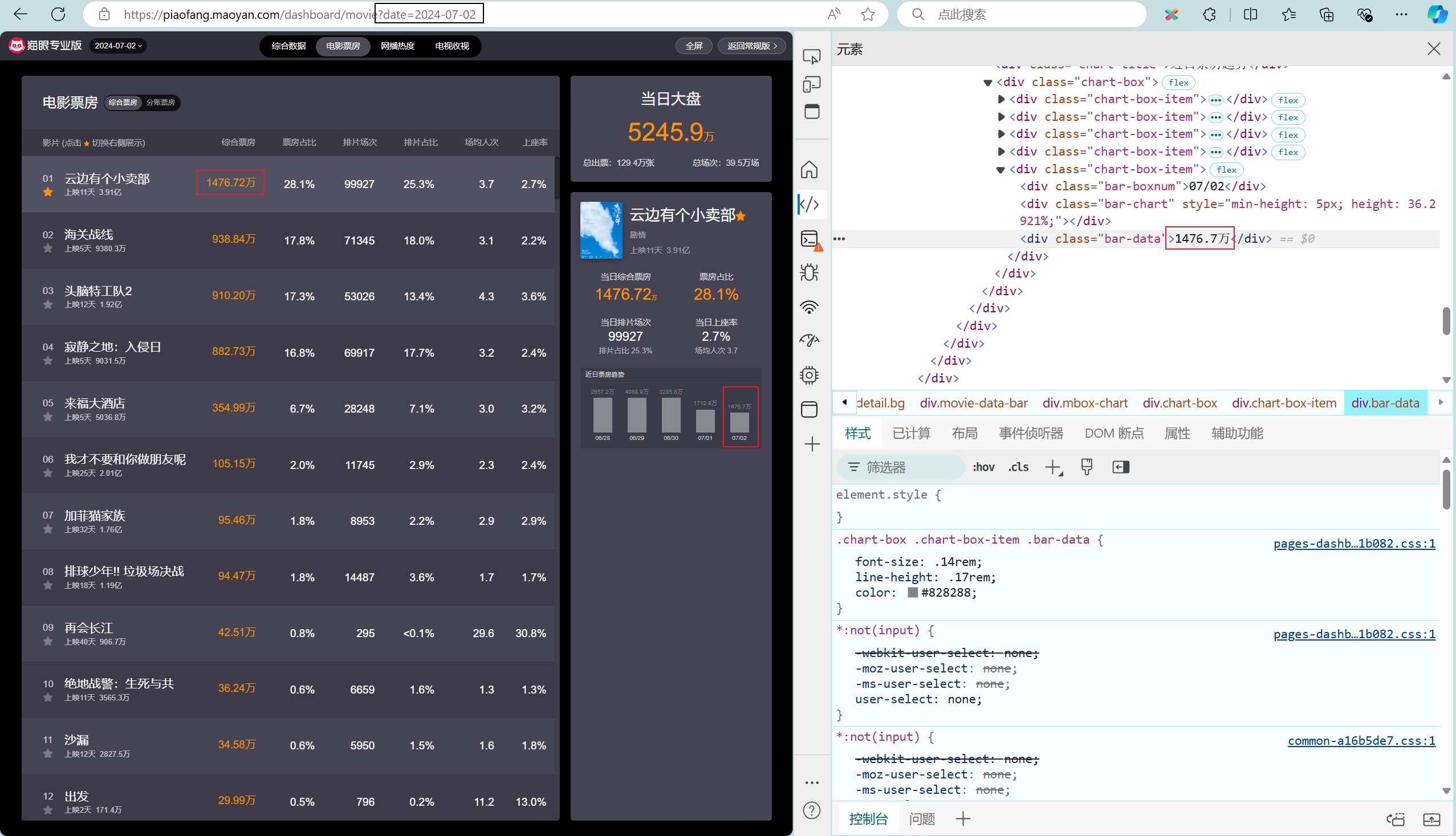
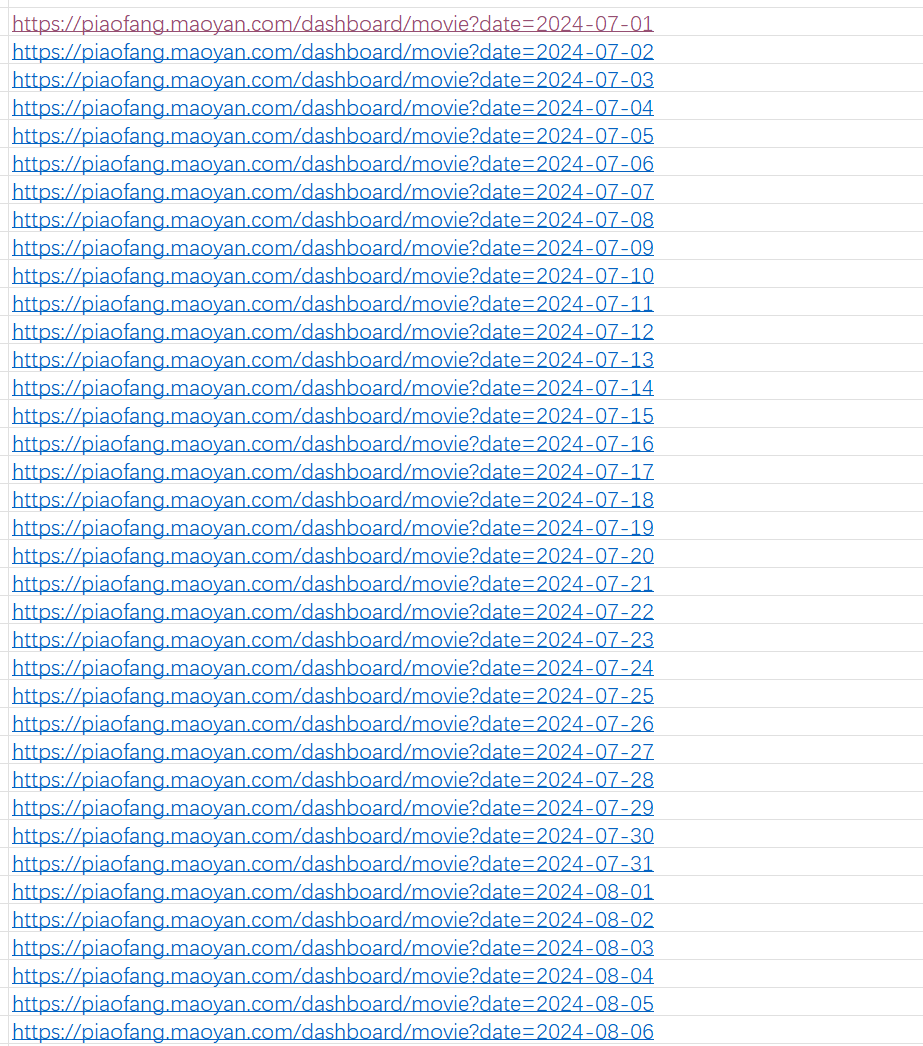
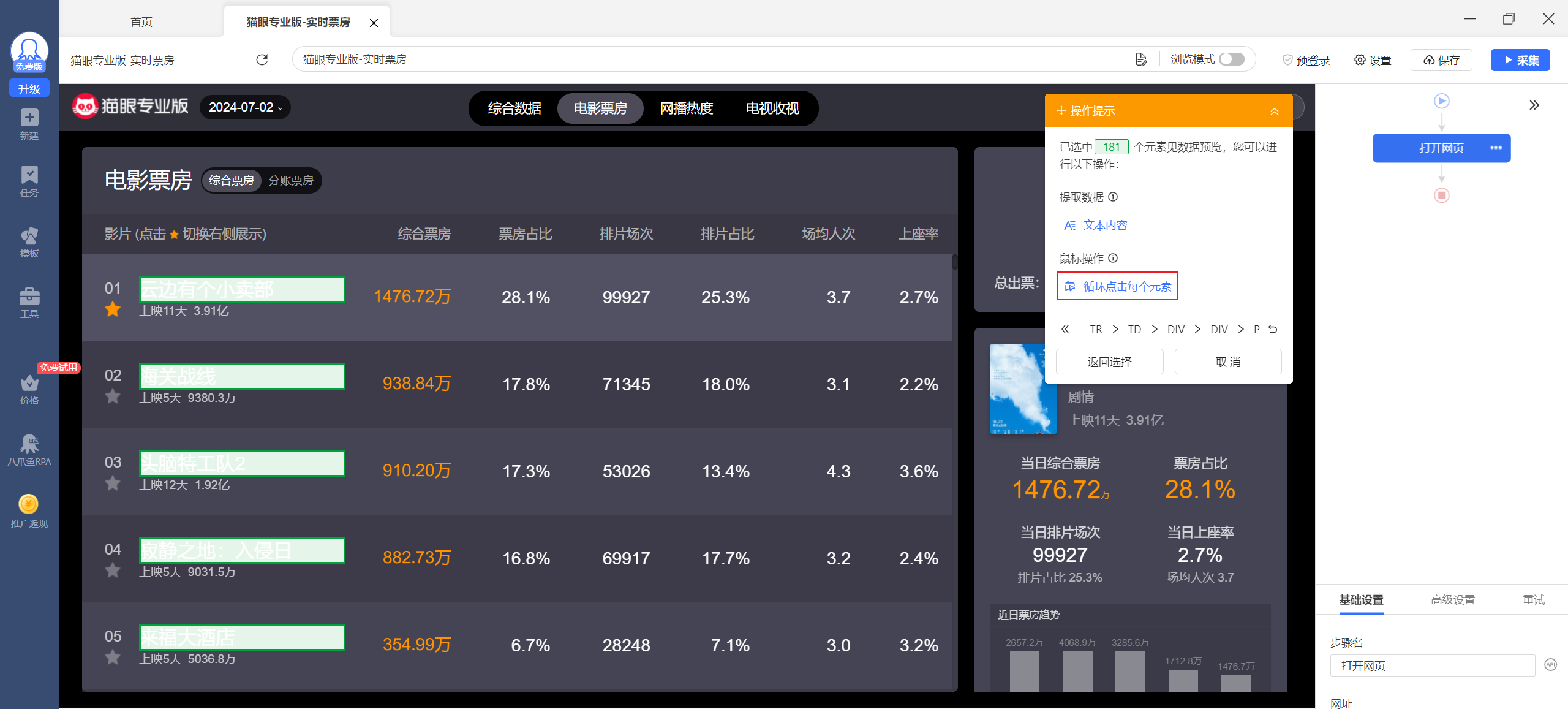
页面登录后，我们选择到2024-07-01，按f12



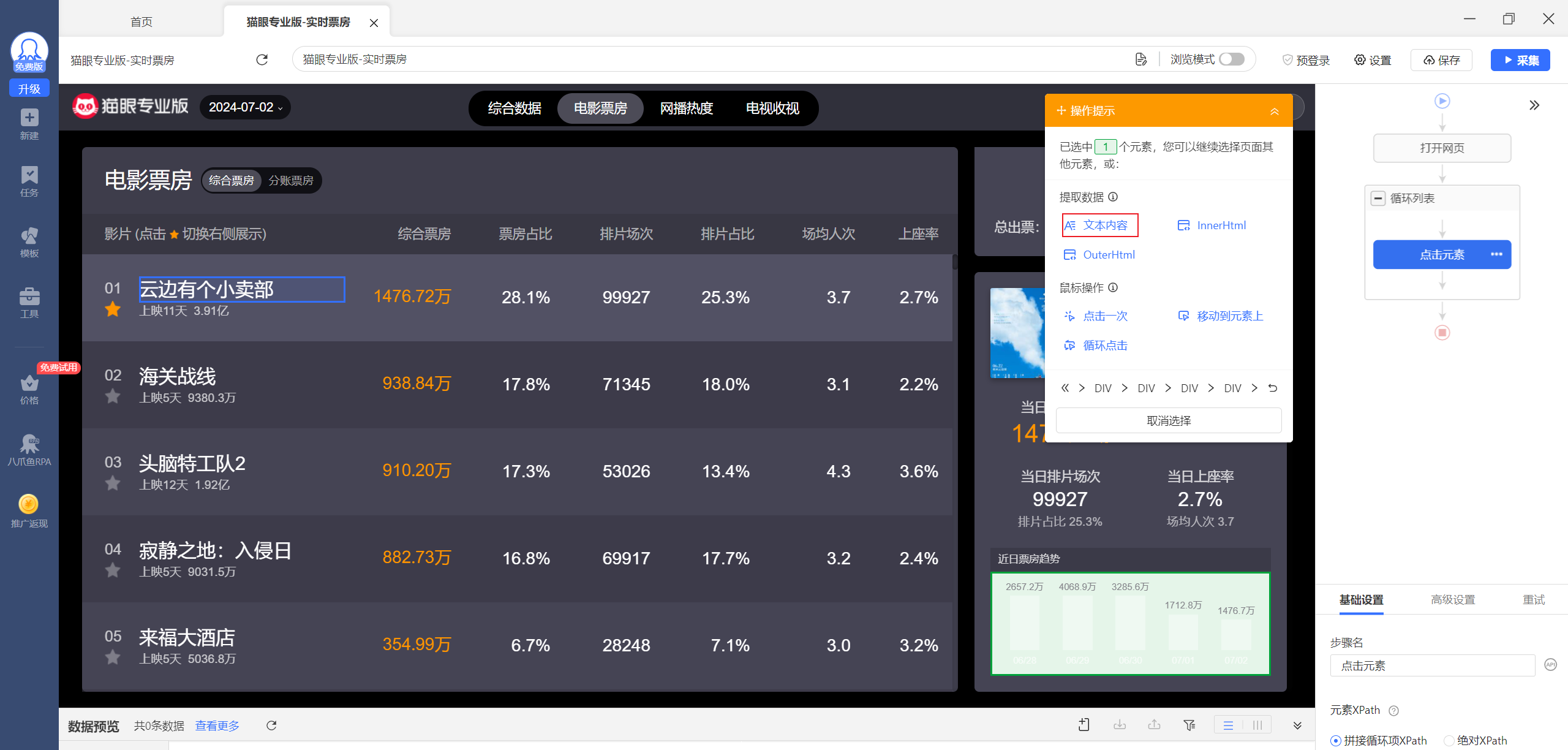
发现，他的票房数据不是数字而是乱码，但是如果点进去一个电影详请，我们可以发现网页右边有一个统计图，图上有近5天的票房数据，而最后一个正是我们想要的。网页URL是属于那种页面传参形式的，?date=后面接的是日期，那如果要提取数据，需要生成一系列date的url，这里使用excel自动填充



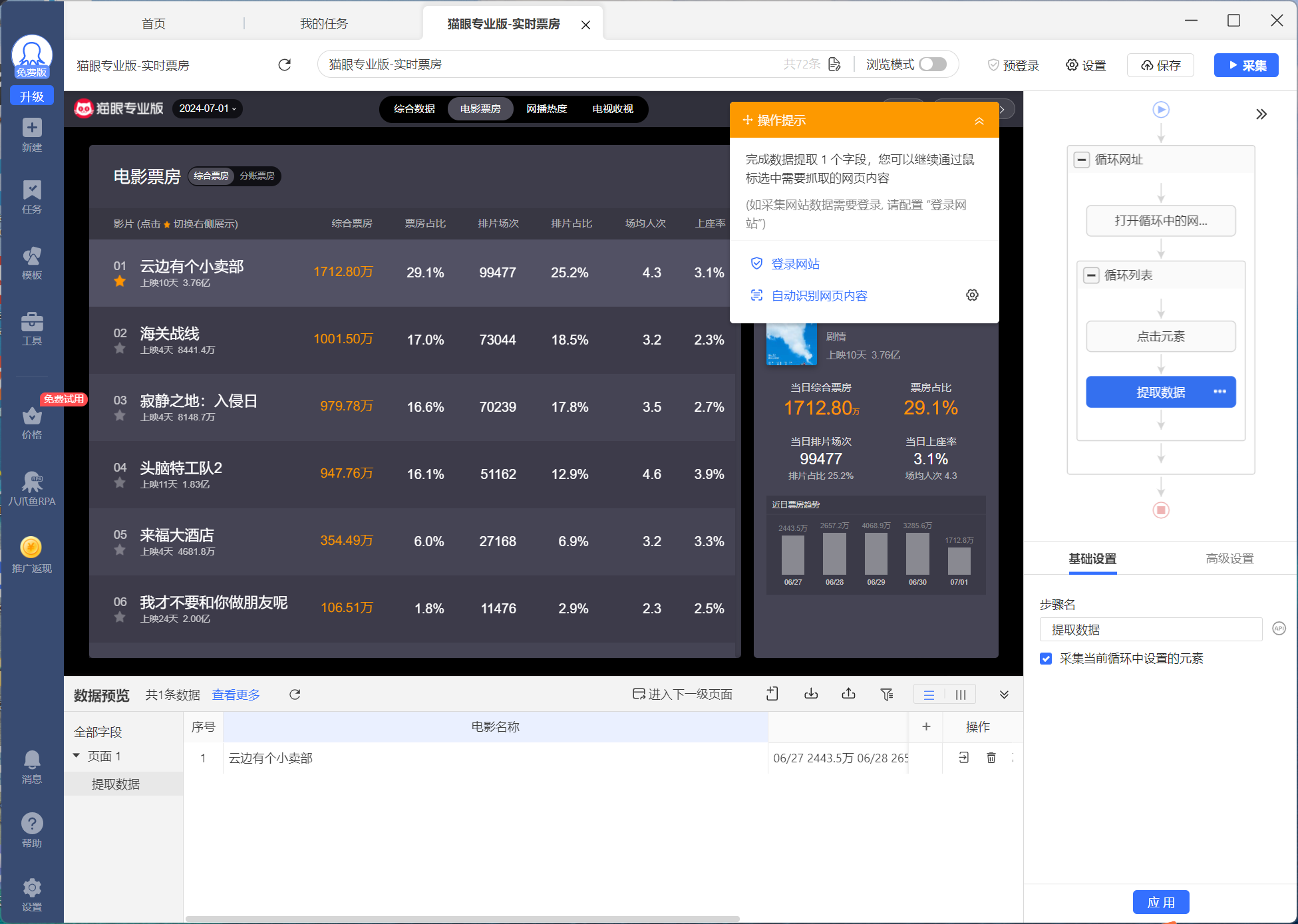
填充完毕后，导入八爪鱼任务，将每一个票房循环点击



随后选中整个图表，提取图表里面的全部内容。这里注意不能直接选择我们想要的今天的数据，**因为某些电影统计了近5天数据，而有些电影仅仅统计了4天数据，不通用，只能全部选择**



添加电影名称，日期等元素



开始采集



随后使用pandas将文本进行简单处理，利用split将文本列分割，提取最后一个字段

import pandas as pd

df = pd.read\_csv('猫眼专业版-实时票房.csv')

df.head()

# 提取提取文本列，按照“\n”分割，提取最后一个和倒数第二个元素

new\_df = pd.DataFrame()

new\_df["电影名"] = df["字段2"]

new\_df['日期'] = df['文本'].str.split('\n').str[-2]

new\_df['日期'] = '2024/' + new\_df['日期']

new\_df['票房'] = df['文本'].str.split('\n').str[-1]

# 若日期列为空或存在汉字，则删除该行

new\_df = new\_df.dropna(subset=['日期'])

new\_df = new\_df[~new\_df['日期'].str.contains('日')]

# 保存结果为excel文件

new\_df.to\_excel('猫眼专业版-实时票房-处理后.xlsx', index=False)