2023 年十部热门电影分析

武汉大学国家网络安全学院

汪宇恒

实验结果简述: 本次实验挑选了 2023 年十部热门电影进行分析, 涉及以下的三个方面:

(1) 每部电影的评分, 即 1 星 ~ 5 星分布比例;

(2) 每部电影的粉丝地区分布属性;

(3) 每部电影的票房属性及十部电影的横向比较;

信息来源: 百度、豆瓣

1 十部电影的评论分析

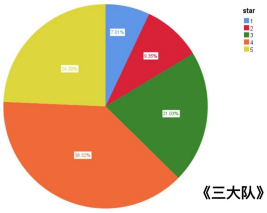
样本概述: 本次实验采集了十部 2023 年的热门电影, 九部电影为国产电影, 一部为热门欧美电影, 且类型涵盖面较广。

选择的电影:《三大队》、《孤注一掷》、《消失的她》、《长安三万里》、《流浪地球 2》、《满江红》、《封神》、《奥本海默》、《八角笼中》、《河边的错误》

1.1 各电影不同评级的分布比例

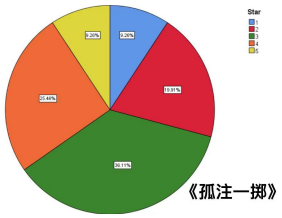
每部电影分别选取了来自豆瓣官网的 200+ 条评论, 存入 excel 表格中, 并利用 SPSS 进行了数据分析, 制作饼状图将数据可视化。

十部电影的评级饼状图如下:



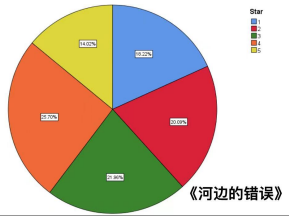
((a)) 《三大队》

[b]



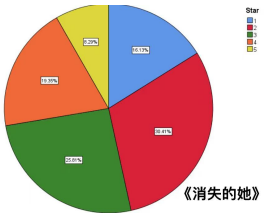
((b)) 《孤注一掷》

[b]



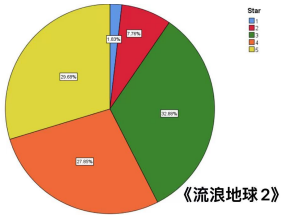
((c)) 《河边的错误》

[b]



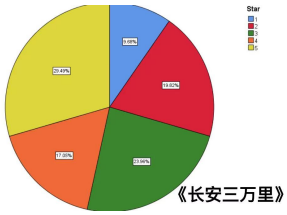
((d)) 《消失的她》

[b]



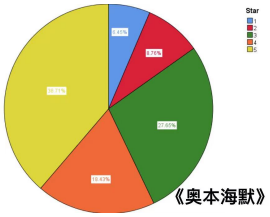
((e)) 《流浪地球 2》

[b]



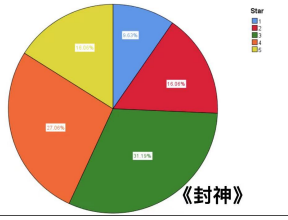
((f)) 《长安三万里》

[b]



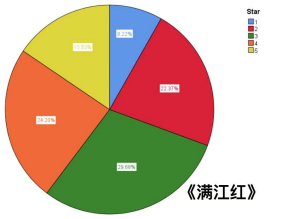
((g)) 《奥本海默》

[b]



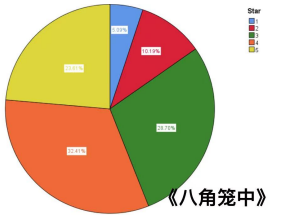
((h)) 《封神》

[b]



((i)) 《满江红》

[b]



((j)) 《八角笼中》

[b]

图 1 十部电影评级频率分布

1.2 各个电影的好评率纵向比较

好评：设置 4 星～5 星为好评

那么，下面开始研究十部电影的好评率纵向比较：

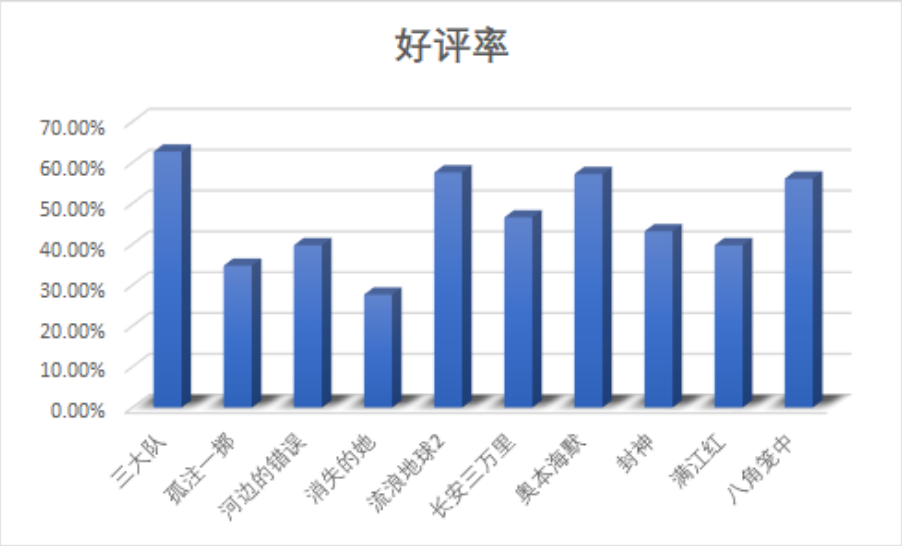


图 2 十部电影好评率比较

结合上图，发现：《三大队》、《流浪地球 2》、《奥本海默》、《八角笼中》四部影片的好评率超过一半，其余六部影片均小于一半。

不难发现，《三大队》是这十部电影中好评率最高的一部电影，而《消失的她》是其中好评率最低的电影。一方面，这个数据可以在一定程度上反映当代影迷对题材新颖程度、生活贴切程度电影的需求；另一方面，也激励着新时代电影产业的不断发展：仅仅依靠改编和俗套的“翻转”，观众是不买账的。

2 十部电影的票房分析

除《三大队》以外，其他电影皆已下映。换言之，只有《三大队》票房会继续增长，剩下的九部电影的票房不会大幅度改变，故可信程度和参考价值较高。

十部电影的票房柱状图如下：

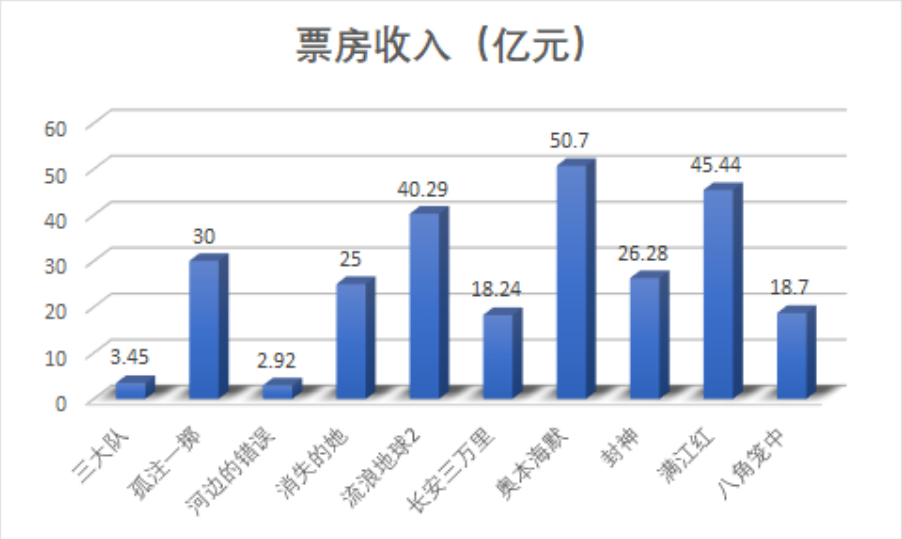


图 3 十部电影票房情况

不难发现，不同电影之间的票房差距是显著的。例如：《奥本海默》的票房足有 50 亿多，而《河边的错误》票房仅仅只有 2.92 亿元。而造成这方面差距的有众多因素，如：受众、影片类型、拍摄投入、上映时间等等。就针对上面的例子而言，《奥本海默》投入成本大，受众较广，且全世界上映，自然能够获得较高的票房。

3 电影的粉丝分布属性

本实验采集了豆瓣网上评论显示的 IP 地址作为相应电影的粉丝所在地址（**注：**每部电影选取的评论数为 230 条），下面就以《消失的她》为例，对其粉丝分布情况进行分析。

安徽	北京	福建	广东	贵州	海南	河北	河南	湖北	湖南	吉林	加拿大	江苏
3	50	4	21	1	1	4	2	11	5	2	3	24
江西	辽宁	美国	内蒙古	山东	陕西	上海	四川	新疆	云南	浙江	中国香港	重庆
3	3	2	2	12	7	21	19	1	4	8	1	3

图 4 《消失的她》影迷分布情况

对上述数据进行处理，分析出《消失的她》各地区粉丝分布数目的一些基础的统计特征：

描述统计

	N	范围	最小值	最大值	均值		标准 偏差	方差	偏度		峰度	
	统计	统计	统计	统计	统计	标准 错误	统计	统计	统计	标准 错误	统计	标准 错误
V2	26	49	1	50	8.35	2.157	10.998	120.955	2.534	.456	7.584	.887
有效个案数（成列）	26											

图 5 《消失的她》粉丝分布统计特征

将数据可视化处理，使《消失的她》影迷分布情况更清晰。

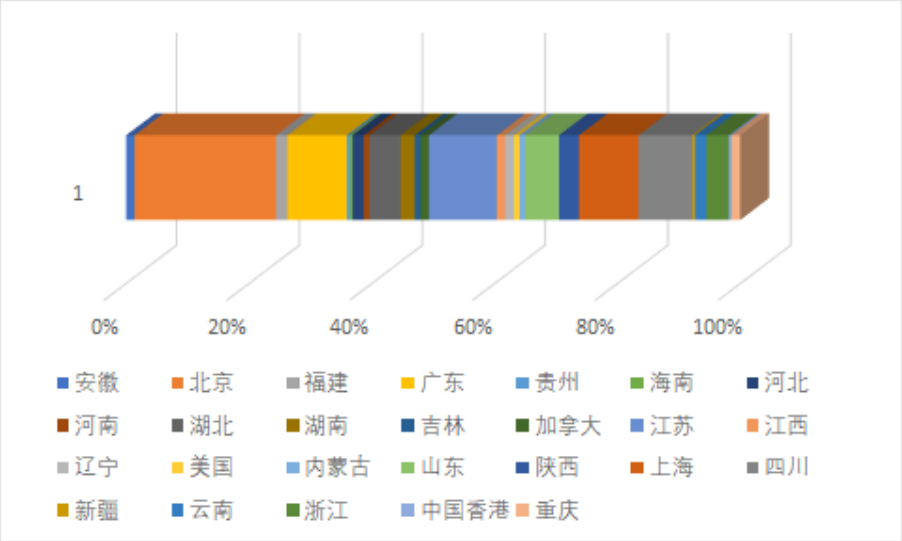


图 6 《消失的她》粉丝分布三维条形图

结合上面图表及数据特征，笔者列出三点最基本信息（可提炼的信息远不止三条）：（1）北京影迷最多；（2）影迷分布范围广泛；（3）各地影迷数量差异较大。

4 数据分析过程

众所周知，评价一步影片的维度是多重的，本次实验从影迷评论、票房收入方面对十部电影进行了分析，下面将展现数据获取以及分析的过程（仅供参考）。

4.1 评论及评级的爬取

为从豆瓣网上获取较为大量的评论及评级数据，本人利用 Python 进行爬虫，将爬的的数据存入 excel 表格中，以供后续的使用以及数据分析。

代码展示如下：

```
import requests
from bs4 import BeautifulSoup
import urllib.parse

import xlwt
import xlrd


def login(username, password):
    url = 'https://accounts.douban.com/j/mobile/login/basic'
    header = {
        'user-agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like
            Gecko) Chrome/73.0.3683.86 Safari/537.36',
        'Referer': 'https://accounts.douban.com/passport/login_popup?login_source=anony',
        'Origin': 'https://accounts.douban.com',
        'content-Type': 'application/x-www-form-urlencoded',
        'x-requested-with': 'XMLHttpRequest',
        'accept': 'application/json',
        'accept-encoding': 'gzip, deflate, br',
        'accept-language': 'zh-CN,zh;q=0.9',
        'connection': 'keep-alive'
        , 'Host': 'accounts.douban.com'
    }

    data = {
```

```
25          'ck' : '',
26          'name': '',
27          'password': '',
28          'remember': 'false',
29          'ticket': ''
30      }
31      data['name'] = username
32      data['password'] = password
33      data = urllib.parse.urlencode(data)
34      print(data)
35      req = requests.post(url, headers=header, data=data, verify=False)
36      cookies = requests.utils.dict_from_cookiejar(req.cookies)
37      print(cookies)
38      return cookies
39
40  def getcomment(cookies, mvid):
41      start = 0
42      w = xlwt.Workbook(encoding='ascii')
43      ws = w.add_sheet('sheet1')
44      index = 1
45      while True:
46
47          header = {
48              'user-agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like
                    Gecko) Chrome/73.0.3683.86 Safari/537.36',
49          }
50
51          try:
52
53              url = 'https://movie.douban.com/subject/' + str(mvid) + '/comments?start=' + str(
```

```
54                start) + '&limit=20&sort=new_score&status=P&comments_only=1'
55            start += 20
56
57            req = requests.get(url, cookies=cookies, headers=header)
58
59            res = req.json()
60            res = res['html']
61            soup = BeautifulSoup(res, 'lxml')
62            node = soup.select('.comment-item')
63            for va in node:
64                name = va.a.get('title')
65                star = va.select_one('.comment-info').select('span')[1].get('class')[0][-2]
66                votes = va.select_one('.votes').text
67                comment = va.select_one('.short').text
68                print(name, star, votes, comment)
69                ws.write(index, 0, index)
70                ws.write(index, 1, name)
71                ws.write(index, 2, star)
72                ws.write(index, 3, votes)
73                ws.write(index, 4, comment)
74                index += 1
75        except Exception as e:
76            print(e)
77            break
78    w.save('test9.xls')
79
80
81 if __name__ == '__main__':
82    username = input('输入账号: ')
83    password = input('输入密码: ')
```

```
84      cookies = login(username, password)
85      mvid = input('电影的id为：')
86      getcomment(cookies, mvid)
```

依靠本程序，实现了将豆瓣网上有关这十部电影的评论均存入了 excel 表格中。(具体代码将在另一文件中展示)

4.2 数据分析以及可视化处理

此处对于数据处理仅进行简要叙述和展示，详细细节不再赘述。

利用的工具：SPSS、Excel

首先先将 Excel 导入 IBM SPSS 中，利用其中的指表工具进行可视化处理。处理结果和过程如下（仅拿《三大队》举例)：

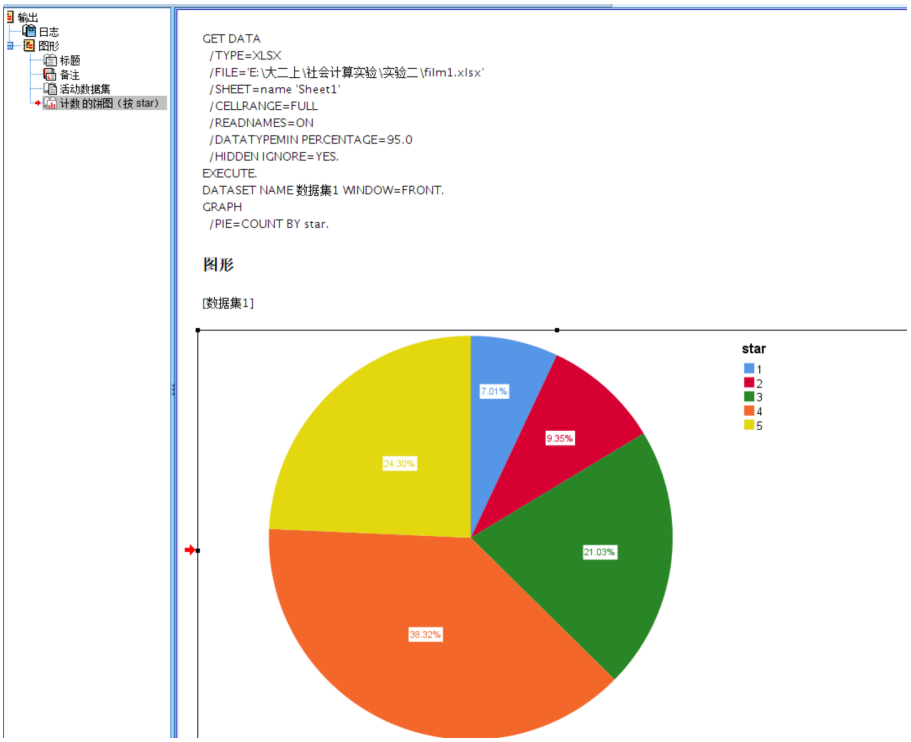


图 7 SPSS 制评级分布饼图

针对票房统计以及各电影的票房和好评率的纵向比较，利用到了 Excel 及其中的制图工具，展示如下：

name	Good_rate
三大队	62.62%
孤注一掷	34.72%
河边的错误	39.72%
消失的她	27.64%
流浪地球2	57.53%
长安三万里	46.54%
奥本海默	57.14%
封神	43.12%
满江红	39.73%
八角笼中	56.02%

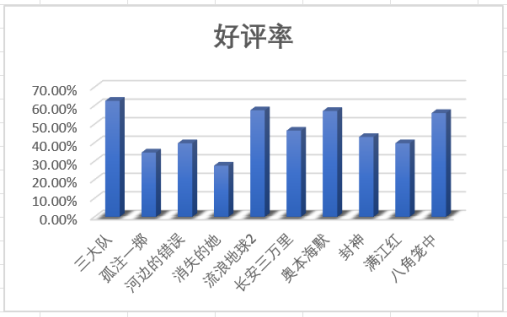


图 8 好评率柱状图制作

name	profit	Good_rate
三大队	3.45	62.62%
孤注一掷	30	34.72%
河边的错误	2.92	39.72%
消失的她	25	27.64%
流浪地球2	40.29	57.53%
长安三万里	18.24	46.54%
奥本海默	50.7	57.14%
封神	26.28	43.12%
满江红	45.44	39.73%
八角笼中	18.7	56.02%

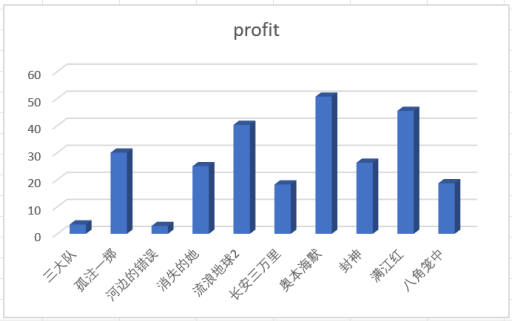


图 9 票房柱状图制作

将好评率和票房制作成柱状图，将数据可视化，可以更清晰地观察出各个电影之间在这两方面的差异。

5 判断票房和好评率是否存在线性关系

除去上面对于电影的分析，笔者还对票房与好评率之间的关系产生兴趣，于是进行了以下实验：

根据朴素的想法，自然总结出：好评率和票房存在一定的正相关：因为好电影必定是口碑好的，而口碑好的属性自然会给其带来高票房的结果。因此，将样本中的十个电影的好评率和票房收入制作成散点图，观察票房和好评率的关系。

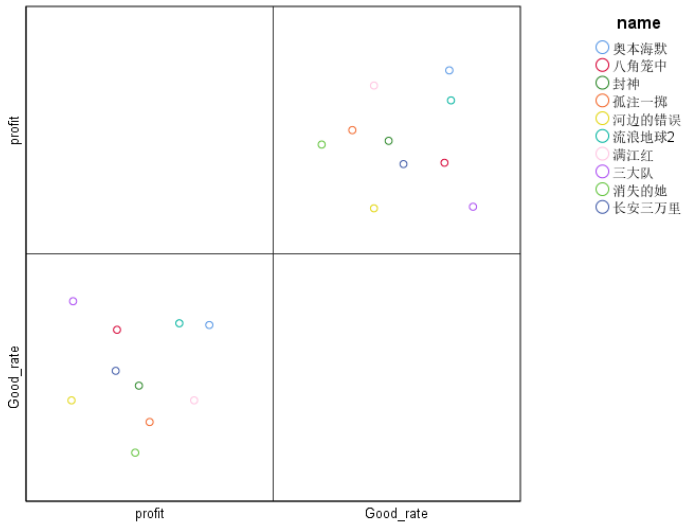


图 10 好评率-票房收入散点图

观察上述散点图，不难发现，票房收入和好评率之间并没有直接的关系，这个结论也可以通过计算皮尔斯相关系数来得到进一步验证。分析产生这个与我们认知相违背的结果的原因，笔者认为有以下三个：（1）样本个数太少，10 部电影不具有代表性，可能当样本个数够大时，二者会呈现出一定的相关性；（2）实际上，大多数影迷并不会上豆瓣评论，故豆瓣上呈现的好评率并不能代表全体影迷的好评率；（3）不排除有部分电影制作商存在着刷票房的行为。

6 总结和 Future work

本次实验，本人学习了运用 Python 爬虫及 SPSS 和 Excel 等工具。学习了处理较大数据量情况下的数据分析。同时，充分利用可视化工具，让数据变得更加清晰、便于分析。切身体会到了“在运用中学习”的真谛。

除去上述研究外，本人对**地区影迷数量与地区 GDP 水平、更大数据量水平下票房和好评率的关系等**研究方向有一定的兴趣，将在未来的工作中体现。

纸上得来终觉浅，绝知此事要躬行。——陆游《冬夜读书示子聿》