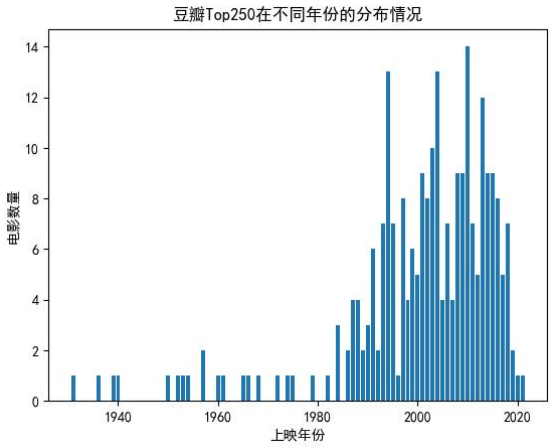
实验报告

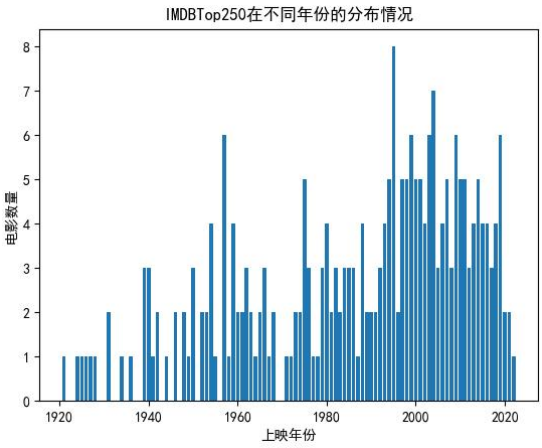
实验内容：
对豆瓣 Top250 与 IMDbTop250 的数据分析与对比。

简介：
豆瓣是国内最大的影迷社区和电影数据库，用户绝大多数为中国影迷；IMDb 是全世界最大的电影数据库，用户来自世界各地，但美国影迷占据了很大一部分。两个榜单都具备着鲜明的特性，因此可以用豆瓣数据来代替中国影迷的观点，用 IMDB 数据代替美国影迷的观点选取这两个网站的数据进行对比的主要目的是来研究中美两国影迷对于电影的不同态度。

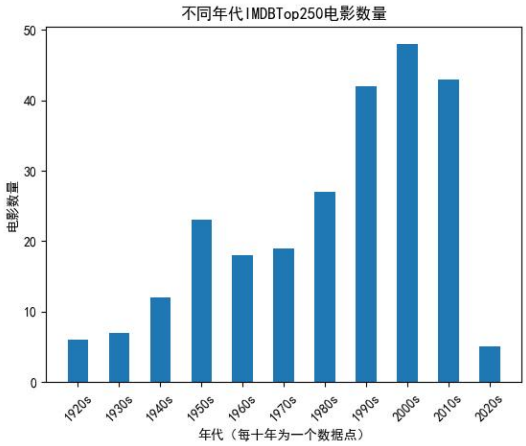
实验过程：
首先分别爬取豆瓣 Top250 和 IMDbTop250 的相关信息（这里对 IMDBTop250 的数据用多种方法均爬取失败，可能是 IMDb 有更加严格的反爬措施，故使用了 Kaggle 上的数据库内的数据来代替）

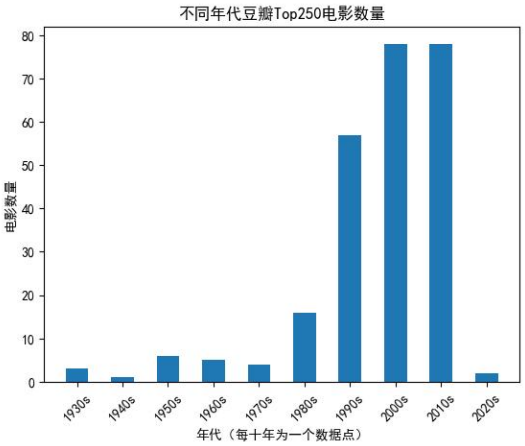
一、Top250 电影的上映年份组成





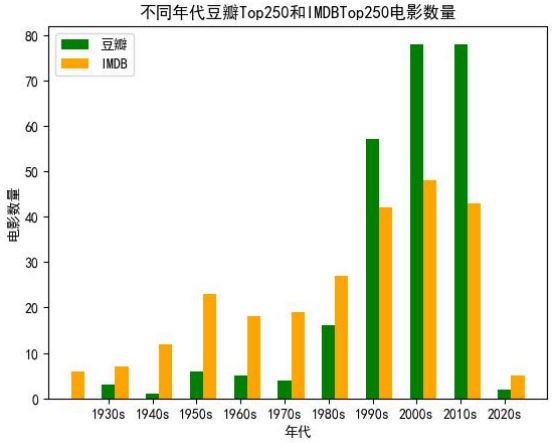
年份数据点太多太乱不够直观，因此对数据点进行整合，以每十年为一个年代。





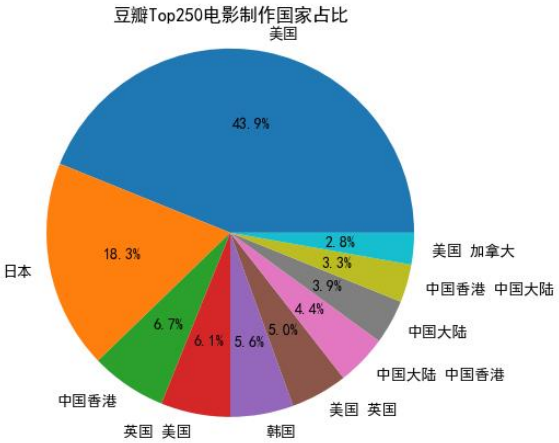
首先对豆瓣的数据进行分析，排名前三的三组数据分别是 00 年代，10 年代和 90 年代，而且三者加起来占据了豆瓣榜单的大约八成，这可能的原因是豆瓣用户群

体中以年轻人居多。而对 IMDb 分析，同样排名前三的数据是 00 年代，10 年代和 90 年代，但这三者加起来只占到 IMDb 榜单的六成左右。

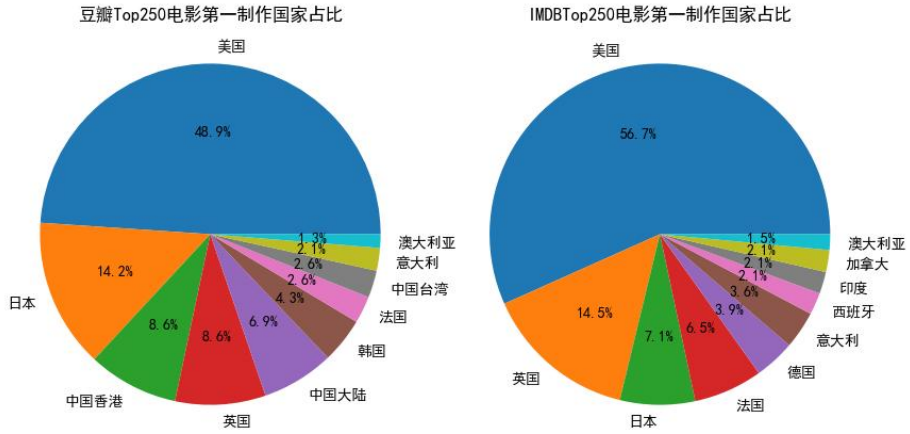


通过分析豆瓣与 IMDb 的数据图，我们发现两者的数据都主要集中在 00 年左右，而对于在 90 年代之前的数据，IMDb 的电影数量明显多于豆瓣的电影数量。我们看到 00 年左右的电影同时收到了两个榜单用户的喜爱，可以推断出从 90 年代到 10 年代这段时间是电影最为辉煌的时候。而对于 90 年代之前的数据，IMDb 数量明显多于豆瓣，可以看出 IMDb 相较于豆瓣更加偏向于老电影。

二、Top250 的制作国家组成



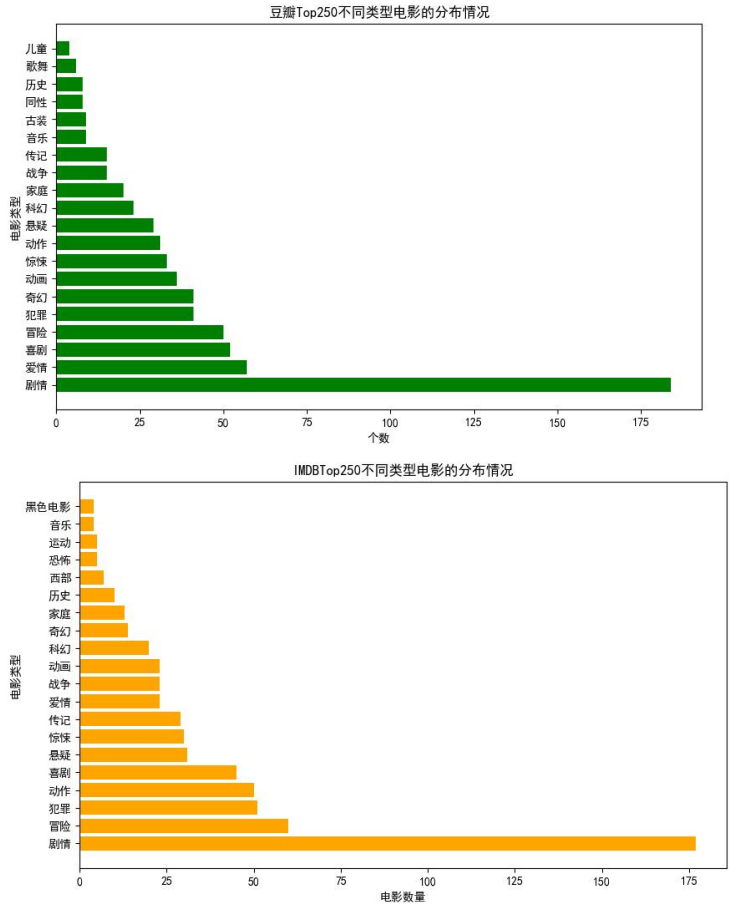
如果按制作国家分类情况太多太杂，所以对制作国家取第一制作国家来进行统计。



通过分析豆瓣与 IMDb 的饼图，我们发现两者的数据占比最多的都是美国的电影，

豆瓣第二多的国家是中国（中国香港、中国台湾和中国大陆加起来的比例），而 IMDb 却是英国，两者的第三名同样是日本。两者排行第一都是美国，足以见得美国电影不仅在自己国家的观众中得到了认可，还赢得到了中国观众的青睐，但在豆瓣排行第二的中国，反而在 IMDB 中排不上号，说明我们的电影在美国的认可度没有那么高，由此可见我国在电影发展方面依旧是道阻且长。豆瓣前十和 IMDb 前十的其他差异，豆瓣前十有韩国，而 IMDb 则有印度、西班牙和意大利，这其中应该是出于文化方面的认同感，中国观众对于亚洲文化认同感较强，而美国观众对于拉美文化更为认同。（这其中印度的数据较高一方面是由于印度本身电影制作较为成熟，另一方面据说是之前存在刷分的现象）

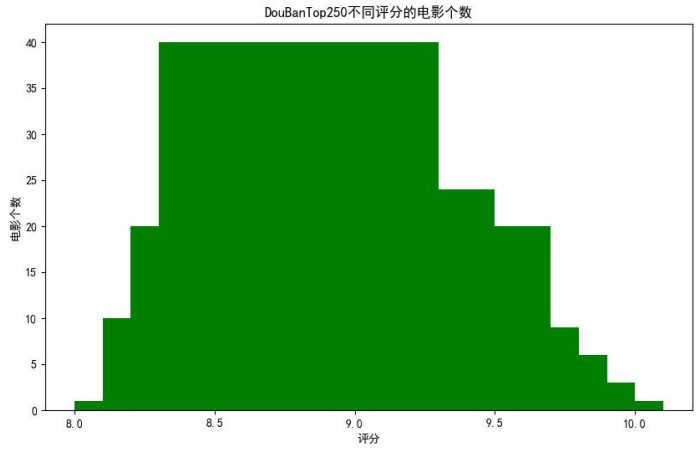
三、Top250 的电影类型组成

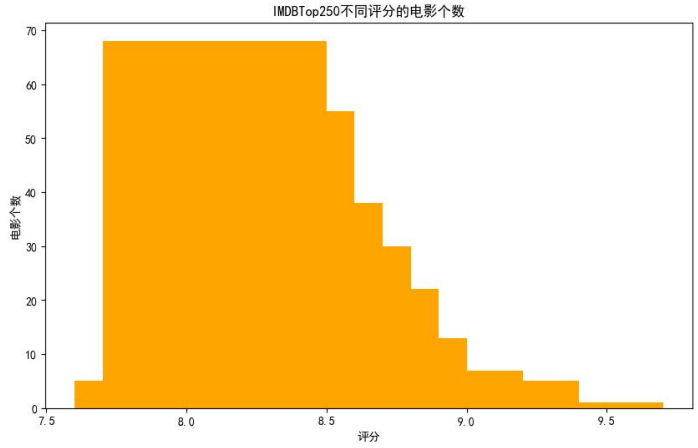


通过柱状图的方式观察并不是很直观，所以采用了词云的方式呈现，字体越大，代表该题材在榜单中的数量越多。

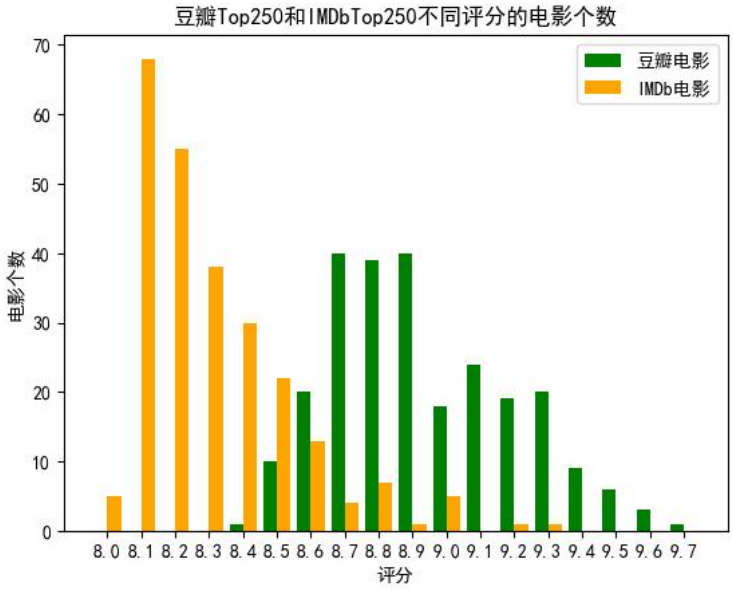
通过对两者电影类型分布数据的分析，我们发现两者都是剧情片占据了绝对的主导地位，豆瓣排名 2、3 的分别为爱情、喜剧，而 IMDb 排名 2、3 的分别为冒险、犯罪，这一点可以看出豆瓣用户更偏向于感性的电影，而 IMDb 用户偏向于刺激的动作电影。而反观当下最热门题材之一的科幻，在两个榜单占据的比重都很少。两者榜单中均有各自独有的题材，豆瓣榜单中有中国独特的古装题材，而 IMDb 榜单中有美国独特的西部题材，这也足以见得两个榜单的鲜明特征。

四、Top250 的评分分布



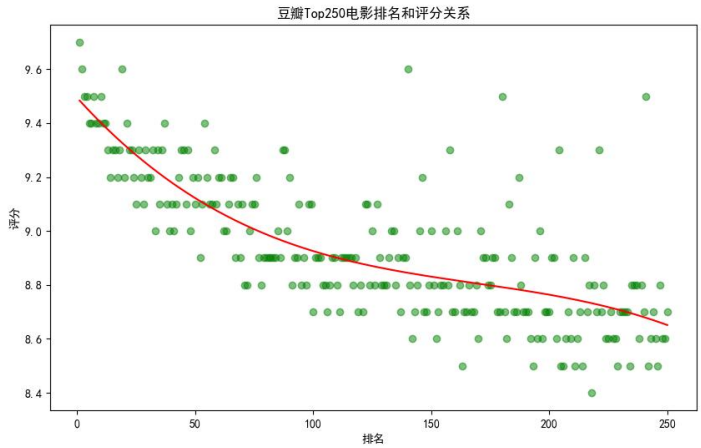


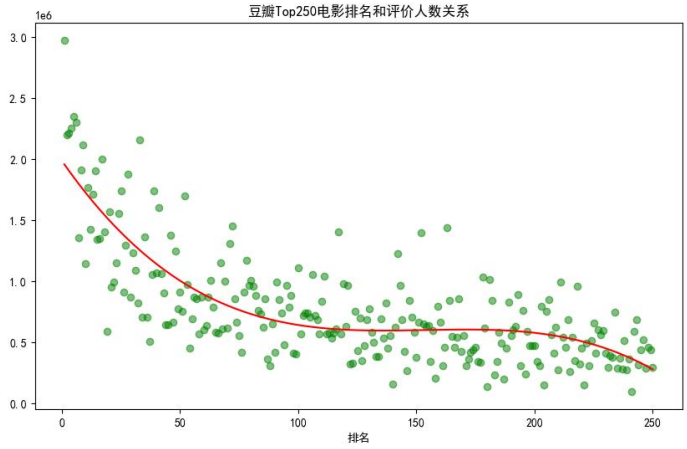
豆瓣的评分主要集中在 8.6 到 9.3 的区间内，最高分为 9.7，最低分为 8.4；IMDb 的评分主要集中在 8.1 到 8.6 的区间内，最高分为 9.3，最低分为 8.0。

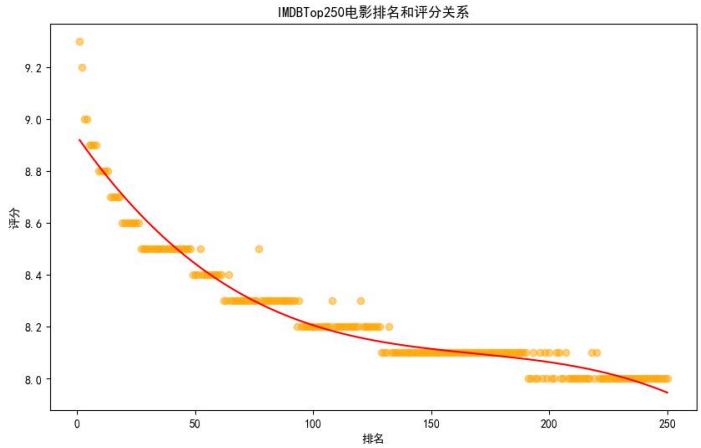


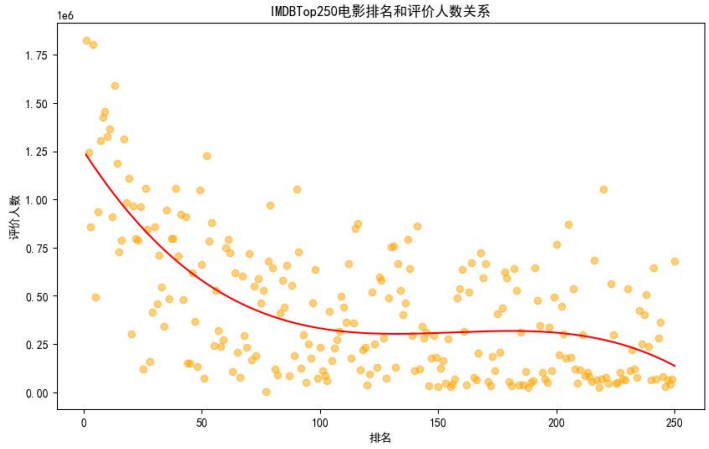
对以上两组数据进行分析，发现豆瓣电影的总体评分比 IMDb 要高，豆瓣的均分为 8.9356，而 IMDb 的均分为 8.3072。我们可以看出豆瓣对于电影较为温和，更倾向于打高分，而 IMDb 对于电影较为苛刻，高分并不多，这可能是 IMDb 用户中专业电影人士占比较大导致的。

五、Top250 的评分和评价人数的分布

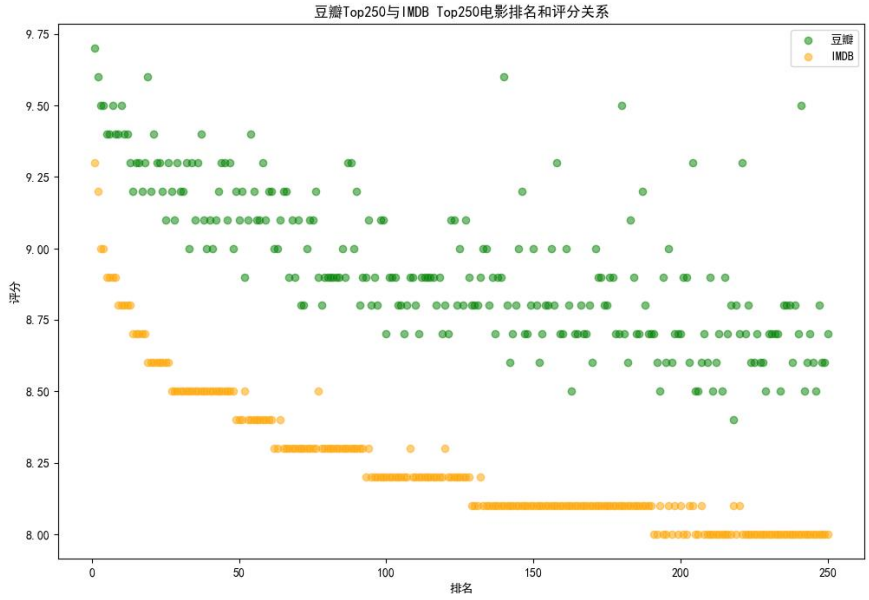

豆瓣Top250电影排名和评分关系
评分
排名


豆瓣Top250电影排名和评价人数关系
1e6
排名

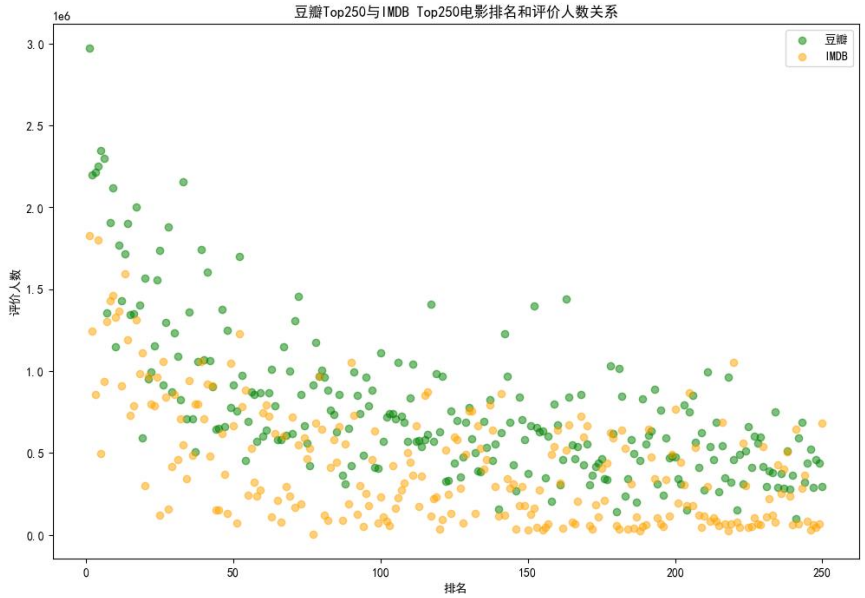

IMDBTop250电影排名和评分关系
评分
排名


IMDBTop250电影排名和评价人数关系
1e6
评价人数
排名

我们对拟合的图像进行分析，发现拟合的函数总体呈现单调递减，即无论是评分还是评分人数，和排名的关系都呈负相关，即评分越高排名越靠前，评分人数越多排名越靠前。



我们能够看出豆瓣评分普遍比 IMDb 高，但 IMDb 的分布更加整齐和集中，方差更小，层次感很强，即高分的排名高，低分的排名低，例外情况很少；而豆瓣的分布更为杂乱和分散，方差更大，很多排名不是那么靠前的电影的评分也很高，基本看不出断层的感觉。

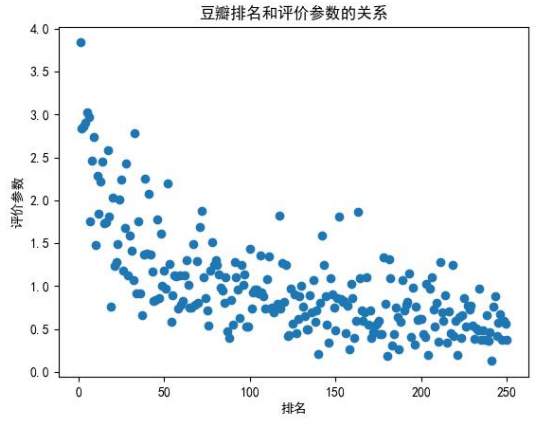


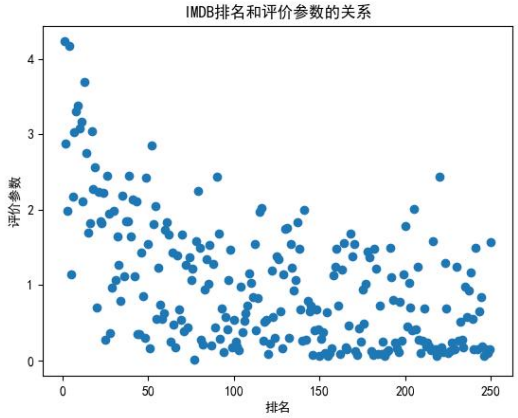
我们可以看出豆瓣评价人数普遍比 IMDb 多，但实际 IMDb 用户应该比豆瓣用户要多，这里可能是由于 IMDb 的评价人数经过筛选了（后面会提到）。

六、对豆瓣和 IMDbTop250 排名标准的简单猜测与模拟

根据一般常识进行初步猜测：影响排名的因素分别为评分和评价人数。然而两者在排名标准中谁占比较多呢?

首先对豆瓣 Top250 进行分析，如果与两者都有关，初步假设为加权平均法。即评价参数=(c)*v/V+(1-c)*s/S 其中 v 为评价人数，V 为平均值，s 为评分，S 为平均值，c 为比例系数。





（详细动态图表参考 python\DaSEIntroduction\final\codes\DouBan\simulation.py 和 python\DaSEIntroduction\final\codes\IMDB\simulation.py）

通过上述实验，我们发现加权平均值的方法和真实情况仍有很大差距。但通过实验，我们可以看出评价参数的影响因素中，应该是评分占据了更大的比重。

通过查找资料，发现豆瓣并未公布其排名机制，而 IMDb 的排名机制采用了贝叶斯统计的算法得出的加权分，公式如下：weighted rank = (v ÷ (v+m)) × R + (m ÷ (v+m)) × C 其中 R 是用普通的方法计算出的平均分，v 为投票人数（需要注意的是，只有经常投票者才会被计算在内，即 IMDb 的投票人数显示的是经常投票者，这样做能够在一定程度上减少刷分行为）m 为进入 IMDb top 250 需要的最小票数，C 为目前所有电影的平均得分。

总结：

豆瓣 Top250 和 IMDBTop250 榜单存在着不少共性，但也有着各自鲜明的特性。豆瓣倾向于 20 世纪之后的电影，而 IMDB 对于 20 世纪之前的老片也同样赞誉有加；豆瓣偏好于温情的题材，而 IMDB 偏向于更加严肃乃至残酷的题材；豆瓣的评分较为宽容，而 IMDB 则较为苛刻。