热门电影影评数据爬取与分析项目分析报告

摘 要

在电影行业飞速发展的当今，电影已经成为了非常普遍的娱乐选择，电影行业无疑已经进入了全民关注，全民讨论时代。但随着大家对电影质量和制作水平的要求越来越高，电影出品方想要斩获高票房，就必须把握好消费者的喜好。

通过Python网络爬虫获取主流电影评分网站豆瓣网上关于《复仇者联盟4》评论用户的基本信息、影评内容等数据，从时间上：用户发表短评数量随日期的变化情况、用户发表短评数量随时刻的变化情况、随日期变化，评分变化情况；城市上：用户常居城市分布情况、不同城市的评分情况；会龄上：发布短评用户的会龄分布、会龄对评分的影响几个方面入手进行分析，得到以下结论：

1. 绝大部分观影用户喜欢在影片上映第一天前去电影院观看，随着时间的流逝，观影人数会越来越少，故电影上映前的充分宣传是增加用户前来观看的几率的有效手段；
2. 电影出品方在电影首映后几小时内应当格外注意豆瓣、淘票票、猫眼等网站的评分与评论，如若口碑不错，可以收集这些资料并及时发布至微博等官方平台，如若口碑不好，则更应当在微博等平台加大宣传力度，引起话题，尝试打一场翻身战；
3. 电影出品方若将影片上映时间改为下午或者晚上，想必能吸引更多用户前来观看；
4. 首映当天的打分会比首映之后的打分高，电影出品方任然应当十分重视首映当天的影片打分情况，在打分结果出来后及时采取对应措施；
5. 经济发达的城市的用户比其他城市的用户更热衷于观看电影的行为，如若想增加电影的票房，吸引经济发达城市的用户来观看比起吸引其他城市的用户来观看的成功几率更大；
6. 非经济发达城市对观影有兴趣的用户在所有用户中占比将近有一半，不容小觑，如若想增加电影的知名度，在这些城市中的宣传也应该到位，不能忽视。
7. 若专门制作针对某地传统特色的电影，不但能够保证原有的经济发达地区的票房，还能够特定地针对某个地方，增加此地的票房，不失为一种好方法；
8. 情怀的确能够拉动票房的增长，因此，制片人可以从情怀入手，制作如《复联》至《复联4》一样的一系列的且时间跨度较大的电影，以此来拉动那个年代的用户观看；
9. 目前影片的制作任应以年轻且有独立经济能力去观影的用户作为主要对象，影片应多加些能够吸引此类人群的元素；
10. 情怀虽吸引用户前来观看，但同时用户对影片的要求也会更高一筹，因此更加要注重情节，注重路人是否看得懂此影片。

关键词： 热门电影影评数据爬取 Python 数据分析 《复仇者联盟4》

**一、问题重述**

**1.1问题背景**

在电影行业飞速发展的当今，电影已经成为了非常普遍的娱乐选择，中国电影最近几年也是突飞猛进，越来越多的人走进电影院。而信息媒体的发展也使得所有人都拥有了随时随地与世界各地的人讨论电影的机会，电影无疑已经进入了全民关注，全民讨论时代。

但是随着越来越多的电影上映，越来越大的阅片量，大家对电影质量和制作水平的要求也越来越高。电影出品方想要斩获高票房，就要把握好消费者的喜好，制作出符合市场期待的电影。

在国内，豆瓣、淘票票、猫眼三家网站都有自己所属平台的评分功能，俨然形成了国内的“电影评分三巨头"。电影评分网站上蕴含了大星的电影评论、排行、评分等信息，是消费者择片困难时的重要参考依据，也是电影出品方了解用户喜好的有效途径。

一份科学的电影消费者数据分析报告，是预测票房的重要依据之- :，也能够帮助出品方规划未来电影拍摄计划，对出品方投资有着重要意义。所以，如何获取、分析消费者喜好信息是电影出品方非常关注的事情。

**1.2提出问题**

为了获取、分析消费者喜好相关信息，本项目将通过Python网络爬虫获取主流电影评分网站豆瓣上关于《复仇者联盟4》评论用户的基本信息、影评内容等数据，并对获取到的影评数据进行简单的文本分析，帮助出品方了解用户偏好。

1. **内容分析**

本次项目分析前期采用附件TASK1，TASK2以及TASK3分别爬取了豆瓣官网上《复仇者联盟4:终极之战》（以下简称《复联4》）中的219条好评数据，218条中评数据以及240条差评数据，合计677条数据。以下将分别从时间、城市及会龄三个方面来对这些数据进行具体的分析。

**(一)时间分析**

**1.1分析用户发表短评数量随日期的变化情况**

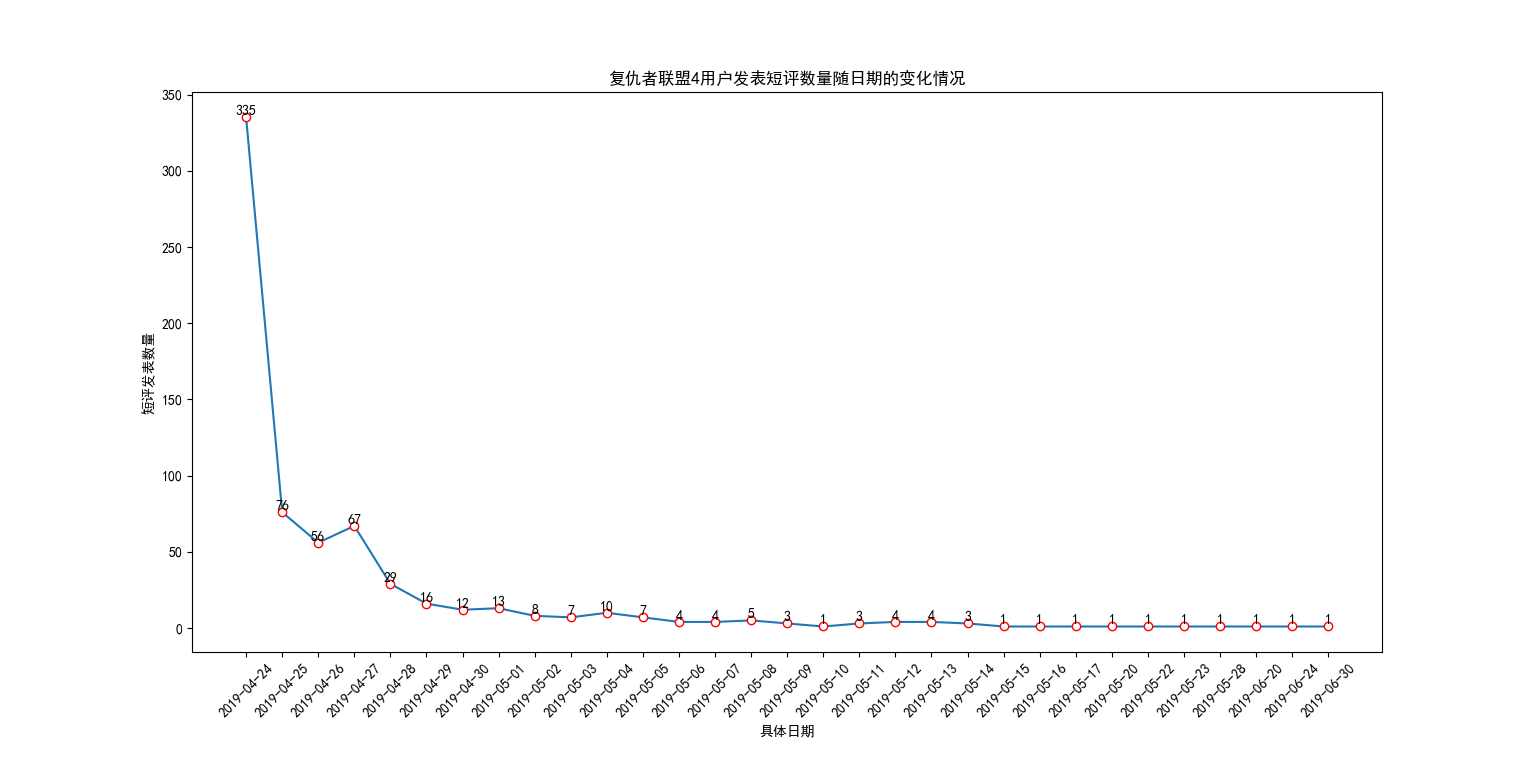
通过附件TASK5，对不同用户发表短评的具体日期进行统计，绘制出《复联4》短评发表数量关于具体日期变化的折线图如图1：

图 1《复联4》短评发表数量随日期的变化情况

从图1中可以看出以下4条特征：

1. 在爬取的677条数据中，有将近一半的用户是在2019年4月24日发表短评的；
2. 在从2019年4月24日开始的为期一周的时间内，用户共发布了591条短评，占据总数据的87.3%；
3. 短评最晚发布时间为2019年6月30日，然而在从2019年5月15至2019年6月30日期间，仅仅有10条短评被发布；
4. 除去2019年4月27日外，图中短评发表数量几乎是随着日期的增加而减少的，呈递减的趋势。

2019年4月24日，这个时间点正是《复联4》在中国大陆的上映时间，将近一半的用户在这一天发表短评的这个现象说明：大部分观影用户喜欢在影片上映第一天前去电影院观看。

而从2019年4月24日开始的为期一周的时间内，用户共发布了591条短评，占据总数据的87.3%，这个现象更是说明绝大部分观影用户都喜欢在影片上映的前期进行观看。

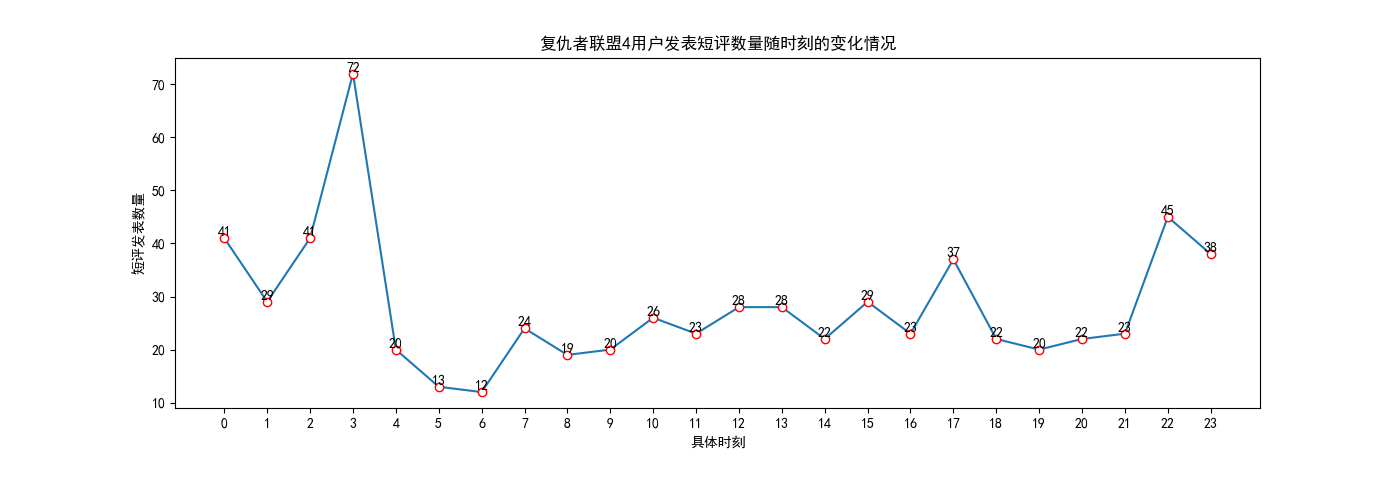
第3、4条特征更是从侧面佐证了上述的结论。随着时间的流逝，观影人数越来越少，半个月之后，观影人数更是低至每天1人甚至更少。

根据上述分析，可以判断出：电影出品方不但应该在电影上映后加紧宣传，更应该注重在电影上映前的宣传。

电影上映前的充分宣传能够勾起观影用户的好奇心，刺激他们选择在电影上映第一天便前来观看。从图中也可以看出，时间越是推移，用户前来观看的几率就越小，因此，电影上映前的充分宣传是增加用户前来观看的几率的有效手段。

**1.2分析用户发表短评数量随时刻的变化情况**

通过附件TASK6，对不同用户发表短评的具体时刻进行统计，绘制出《复联4》短评发表数量关于具体时刻变化的折线图如图2：

图 2《复联4》短评发表数量随时刻的变化情况

从图2中可以看出以下3条特征：

1. 在凌晨3点发表的短评数量比其他时刻短评数量显著地多，是其他时刻短评发表数量的两倍、三倍乃至六倍；
2. 除去凌晨3点，其他时刻短评发表数量较多的有0点、2点、17点、22点及23点；
3. 凌晨5、6点发表的短评虽然有，但是与其他时刻发表的短评数量相比，实在是少。

对于第一条特征，影片《复联4》全长181分钟，约3个小时，其在中国大陆的首映时间为2019-04-24 00:00，故图1中在2019-04-24 00:00进行观看的335人将会在当天凌晨3点01分结束影片的观看，并将随后把自己对观看《复联4》首映无法抑制的感想在豆瓣上以短评的方式展现出来，这是造成图2中凌晨3点发表的短评数量远远高于其他时刻的原因。

而这时的打分与短评数据会成为第二天未观看首映的用户的参考资料，也会成为各家自媒体大肆宣传的材料之一。因此，电影出品方在电影首映后几小时内应当格外注意豆瓣、淘票票、猫眼等网站的评分与评论。

如若口碑不错，可以收集这些资料并及时发布至微博等官方平台，让未观看用户怀有已经看过的人大部分都觉得这部电影很不错，值得一看的想法，从而吸引他们前来观看。如若口碑不好，则更应当在微博等平台加大宣传力度，引起话题，在不妙的评论蔓延之前及时激起未观看影片且未观看评论的用户的好奇心，吸引他们前来观看，进而尝试打一场翻身战。

对于第二条特征，可以发现0点、2点、22点及23点都处于深夜与凌晨的时候，17点是下午吃晚饭的时候，即在中午发表的短评数量并不多。对于一般情况来讲，即使看完电影心情再澎湃，用户也不会选择站在电影院门口发表短评，而是在一个较为空闲的时候（如吃饭前后），且较为方便的地点（如家中、奶茶店、餐馆等）来发表短评，然而观影过后激情总是会消退的，因此想要发表短评的用户也不会等观影后过很久的时间再去发表。

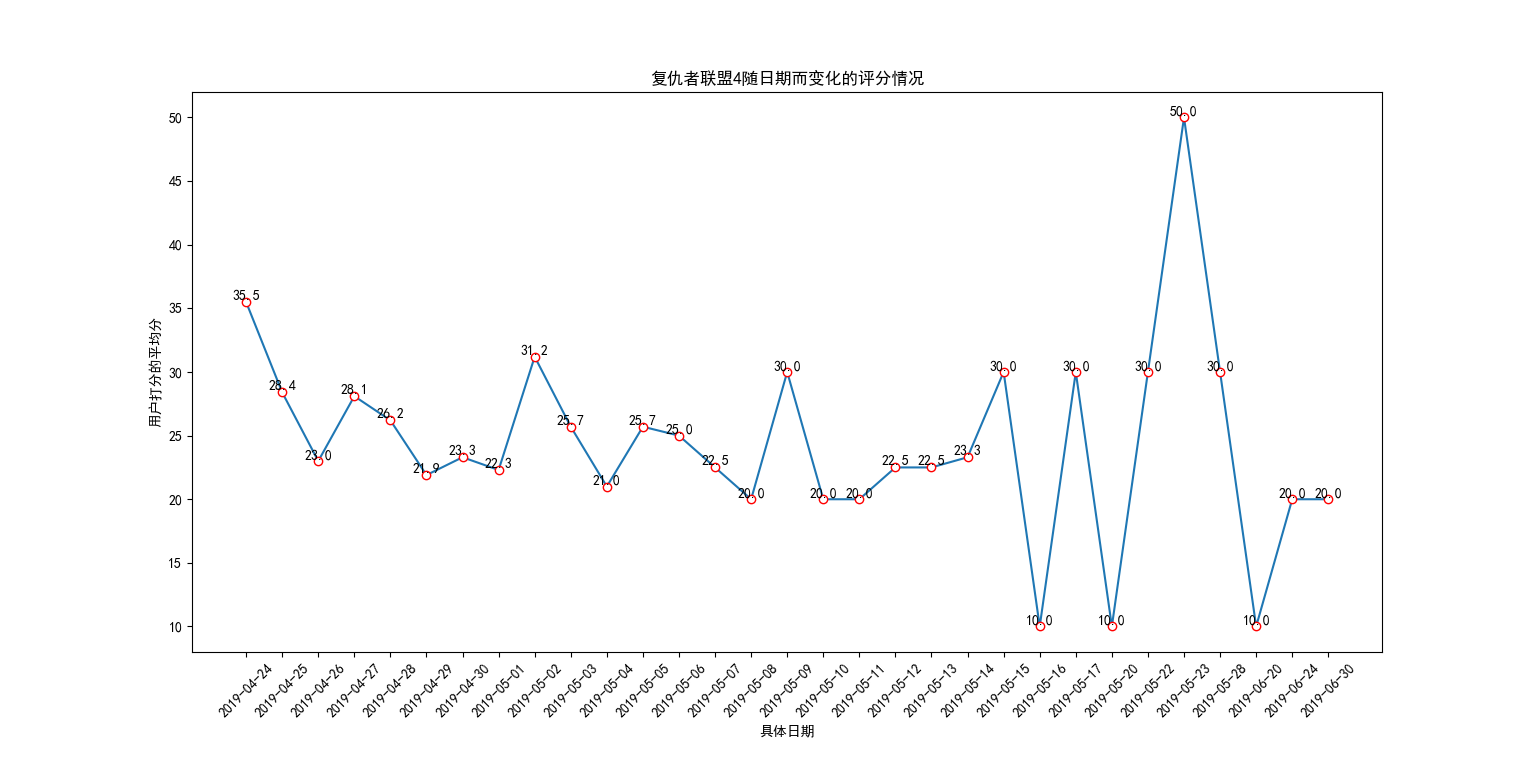
由此可以推测：若是上午观看了电影，则应当在中午的时间发表短评，若是下午观看了电影，则应当在晚饭的时间发表短评，若是晚上乃至深夜观看了电影，则应当在回到家中或找到地方可以坐下来休息时进行发表。而图2中显然短评发表的高峰期不在中午，这说明大部分用户更喜欢在下午或者晚上的时间去观看电影。

故电影出品方若将影片上映时间改为下午或者晚上，想必能吸引更多用户前来观看。

然而由特征3可以看出在凌晨观看影片的情况虽然有，但还是比较少的，因此不建议电影出品方若将影片上映时间定为凌晨。

**1.3分析随日期变化，评分变化情况**

通过附件TASK8，对不同日期用户对影片打分的平均分进行统计，绘制出《复联4》随着日期变化而变化的评分情况如图3：

图 3《复联4》用户打分平均分随日期的变化情况

从图3中可以看出以下3条特征：

1. 从2019年5月15日开始评分变化幅度非常大；
2. 图3中的评分基本围绕在20与35之间，波动不大，较为均衡；
3. 除去2019年5月23日的50分，其余分数中，2019年4月24日当天打分为最高，达到35.5分。

对于第一条特征，由图1已经看到：2019年5月15日之后，每天最多只有一个人发表短评进行打分。因此2019年5月15日之后的打分带有十分浓重的主观思想，不能用于代表大部分用户的想法，因此在分析用户打分平均分随日期的变化情况时，2019年5月15日之后的数据可以舍弃。

对于第二条特征，豆瓣上《复联4》的打分达到了较高的8.6，那么为什么图3中的数据却在较为一般的20与35之间波动呢？在上文已经提到过：本文所提取的数据有219条好评，218条中评以及240条差评，合计677条数据。各种评分所取的数据量大抵相同，这就造成了图3中评分大多数处于中间位置的现象，然而事实上，豆瓣上给《复联4》打40,50分的人远多于给它打30,20,10分的人，因而在豆瓣上《复联4》的打分是较高的8.6分。

但是这并不能说明这组数据是无用的，因为每个人对不同事物的看法不同，不管是给《复联4》打高分的用户，还是给《复联4》打低分的用户，他们都是电影出品方的潜在用户，因此对于这些用户，我们不应该区别对待，而应该一视同仁都进行分析。

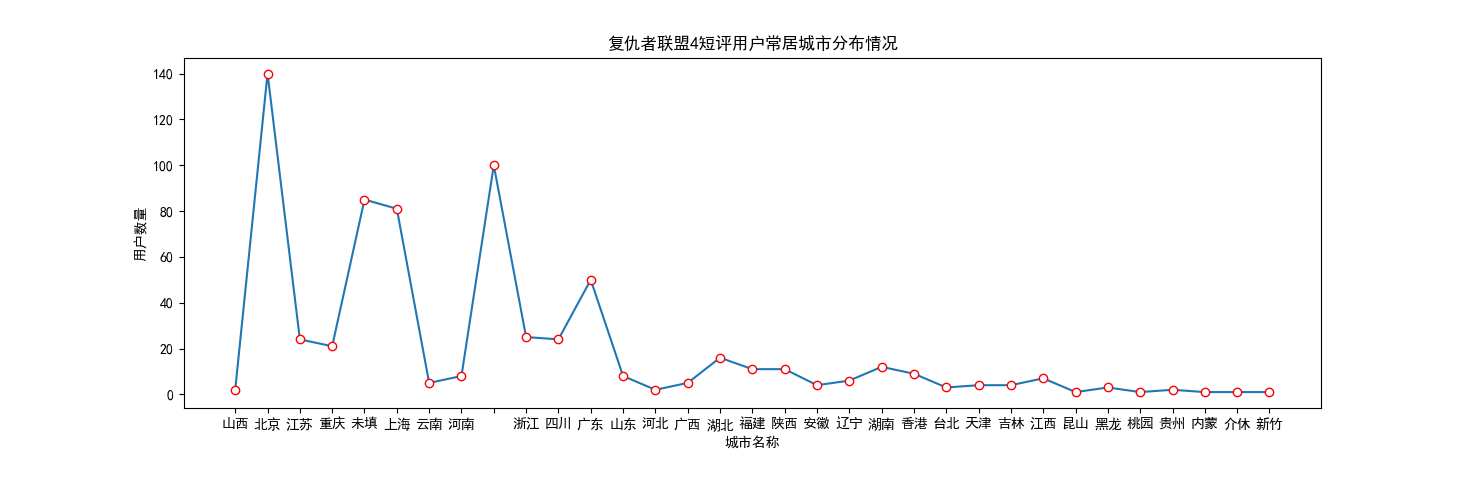
对于第三条特征，可以看出即使好评中评差评总数量差不多，但是首映当天的打分依旧会比首映之后的打分高，因此，电影出品方任然应当十分重视首映当天的影片打分情况，在打分结果出来后及时采取对应措施在盛况时添一把火，吸引未观看的用户前来观看，促使已观看的用户二刷，也及时采取对应措施在低谷时挽回局面。

**(二)城市分析**

每个城市都有自己独特的人文氛围，思想情怀等，因此不同城市观影用户的喜好也不一样，下面将从两个方面：用户常居城市分布情况以及不同城市的评分情况，来分别分析不同常居城市的用户偏好是什么样子的。

**2.1分析用户常居城市分布情况**

通过附件TASK9，对不同用户常居城市的分布情况进行统计，绘制出《复联4》短评用户常居城市分布情况折线图如图4：

图 4《复联4》短评用户常居城市分布情况折线图

数据解释：图中横坐标为城市名称，纵坐标为所爬取用户信息中常居城市为此城市的用户的数量。在横坐标中有两处比较特殊，即“未填”以及一个空的城市名称。“未填”指的是该用户没有在豆瓣上填写他的常居城市，以下将不对此数据进行分析。空的城市名称指的是该用户多所填常居城市在中国范围之外，虽然其数量将近100，但是其较为分散，故将其整合至一起，且不进行分析。

此外，为了对各城市用户数量占比看的更加直观，绘制出《复联4》短评用户常居城市分布情况圆饼图如图5：

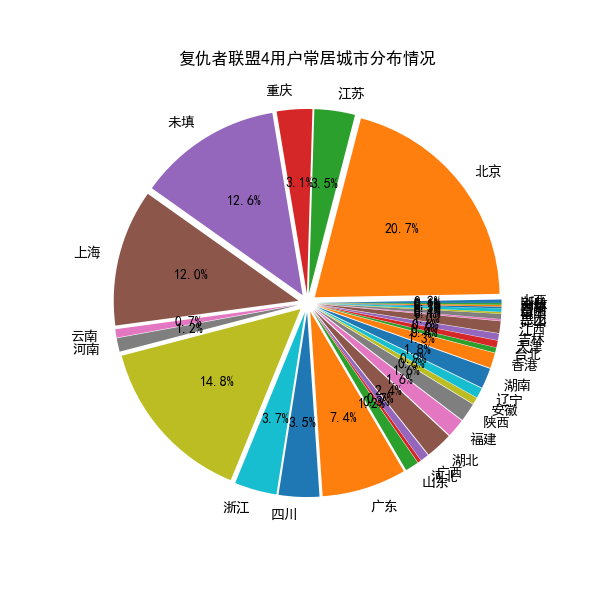


图 5《复联4》短评用户常居城市分布情况圆饼图

从图4与图5中可以看出以下2条特征：

1. 用户常居城市差别较大，其主要集中在北京，上海，广东三个地方，其中北京更是达到140人次，占总数量的20.7%；
2. 除北京，上海，广东，未填，空白这些城市外，其他城市虽然每个占比都不大，仅有1%，2%左右，但其总和达到了所有城市中的48.12%。

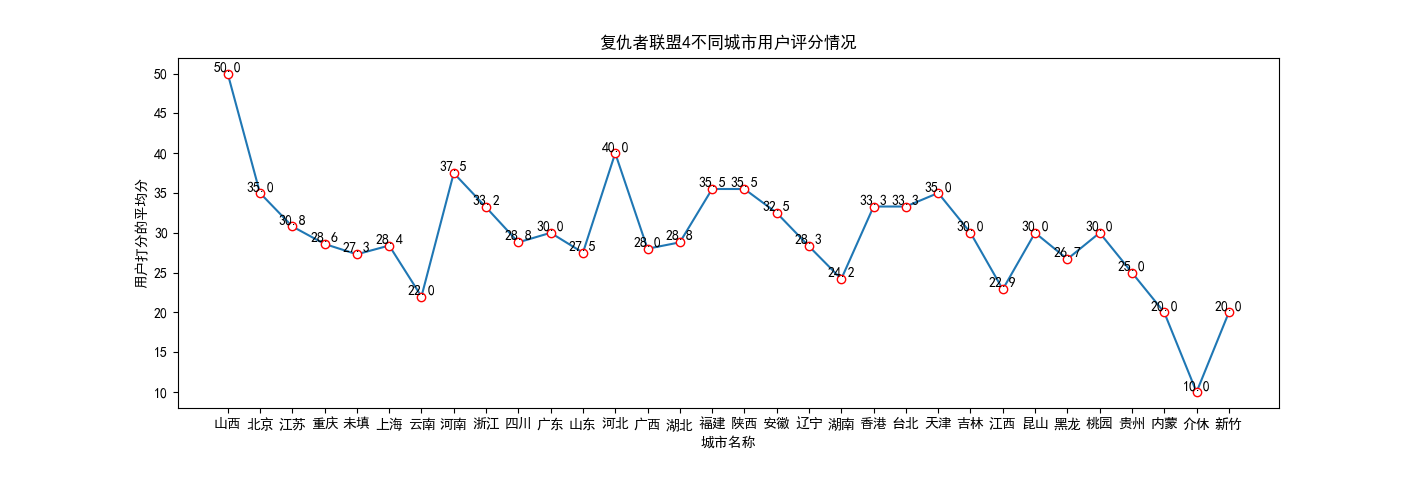
对于第一条特征，北京，上海，广东三个地方都是经济比较发达的城市，可以看出经济发达的城市的用户比其他城市的用户更热衷于观看电影的行为，因此，如若想增加电影的票房，吸引经济发达城市的用户来观看比起吸引其他城市的用户来观看的成功几率更大，为了达到目的，电影出品方可以在影片中加入一些经济发达城市的特色来作为对应手段。

而对于第二条特征，可以看出非经济发达城市也是有部分用户是有着观影的兴趣的，并且这部分用户在所有用户中占比将近有一半，是个不容小觑的用户量。因此，如若想增加电影的知名度，打响名声，那么在这些城市中的宣传也应该到位，不能忽视。

除此之外，若专门制作针对某地传统特色的电影，如景德镇陶瓷，那么首先经济发达地区的用户依旧会前来观看，除此之外，景德镇的用户也会产生亲切感，从而前来观看，此手法不但能够保证原有的经济发达地区的票房，还能够特定地针对某个地方，增加此地的票房，不失为一种好方法。

**2.2分析不同城市的评分情况**

通过附件TASK10，对不同城市用户对影片打分的平均分进行统计，绘制出《复联4》短评用户常居城市分布情况折线图如图6：

图 6《复联4》不同城市用户评分情况

从图6中可以看出以下2条特征：

1. 偶尔有个别较为突出的分数，如山西的50分和介休的10分；
2. 与图3一样总体分数较为均衡，大多数位于25和37.5之间；

对于第一条特征，由于在获取的数据中，多个城市仅有一位用户，如山西与介休，因此这些城市的用户打分也是带有浓重的主观看法的，不能代表这个城市整体对这部影片的打分，故不予分析。

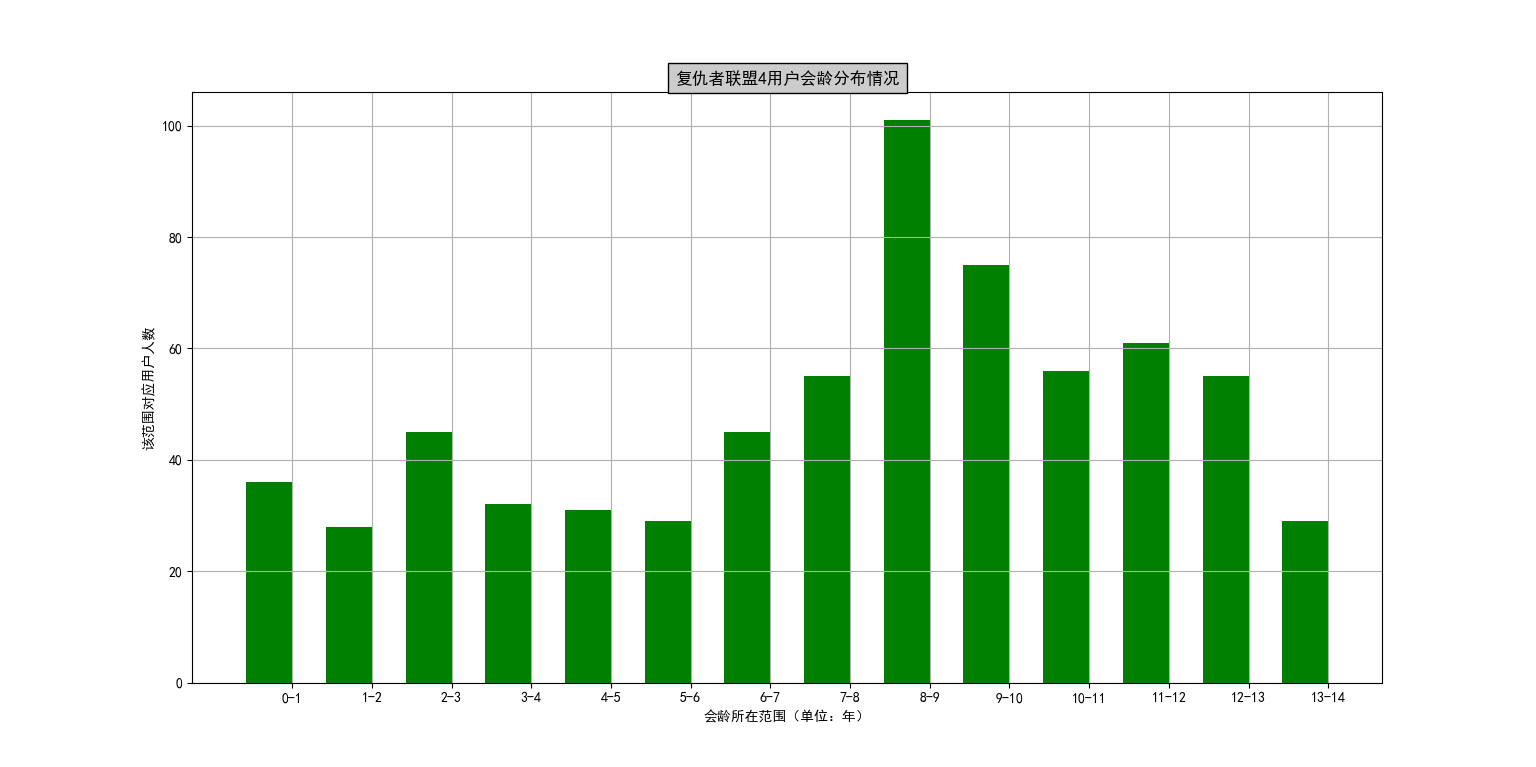
对于第二条特征，上文已经解释过：数据量大抵相同的219条好评，218条中评以及240条差评，合计677条数据便是产生这种现象的原因。

**(三)用户会龄分析**

用户会龄一般与用户年龄挂钩，从用户入会时长可以分析不同年龄段用户的偏好，同时可以从侧面反映该影片的受众群体，从而为不同影片制定营销策略，故接下来将分别从《复联4》发布短评用户的会龄分布以及会龄对评分的影响两个方面来进行分析。

**3.1分析《复仇者联盟4》发布短评用户的会龄分布**

通过附件TASK12，对不同用户的会龄分布情况进行统计，绘制出《复联4》短评用户会龄分布情况柱状图如图7：

图 7《复联4》不同用户会龄分布情况柱状图

从图7中可以看出以下2条特征：

1. 用户会龄分布最多的是8-9年，也就是从2010年到2011年注册的用户；
2. 会龄在8-9年之间的用户数量往左右两边都呈下降趋势，直到4-5年及4-5年之内这个会龄范围，用户数量才较为均衡；

对于第一条特征，第一部《复联》拍摄于2010年，或许正是这个原因，才让会龄在8-9年之间的用户情怀最高，最为期待观看此部影片。这个现象说明情怀的确能够拉动票房的增长，因此，制片人可以从情怀入手，制作如《复联》至《复联4》一样的一系列的且时间跨度较大的电影，以此来拉动那个年代的用户观看。

对于第二条特征，会龄在8-9年之前的，越往前其用户真实年龄就越大，因而对于影片的热情反而会随着年龄的增加而减少，故其数量呈下降趋势。而会龄在8-9年之后的，其真实年龄还小，对于此类影片不能完全理解，故其对于影片的热情反而会随着年龄的减小而减少。

此现象说明影片观看人数目前主要还是以年轻且有独立经济能力去观影的用户为主，其他年代的观影用户虽然有，数量却不能与年轻且有独立经济能力的用户相比，故目前影片的制作任应以此类人群为主，影片应多加些能够吸引此类人群的元素。

**3.2分析会龄对评分的影响**

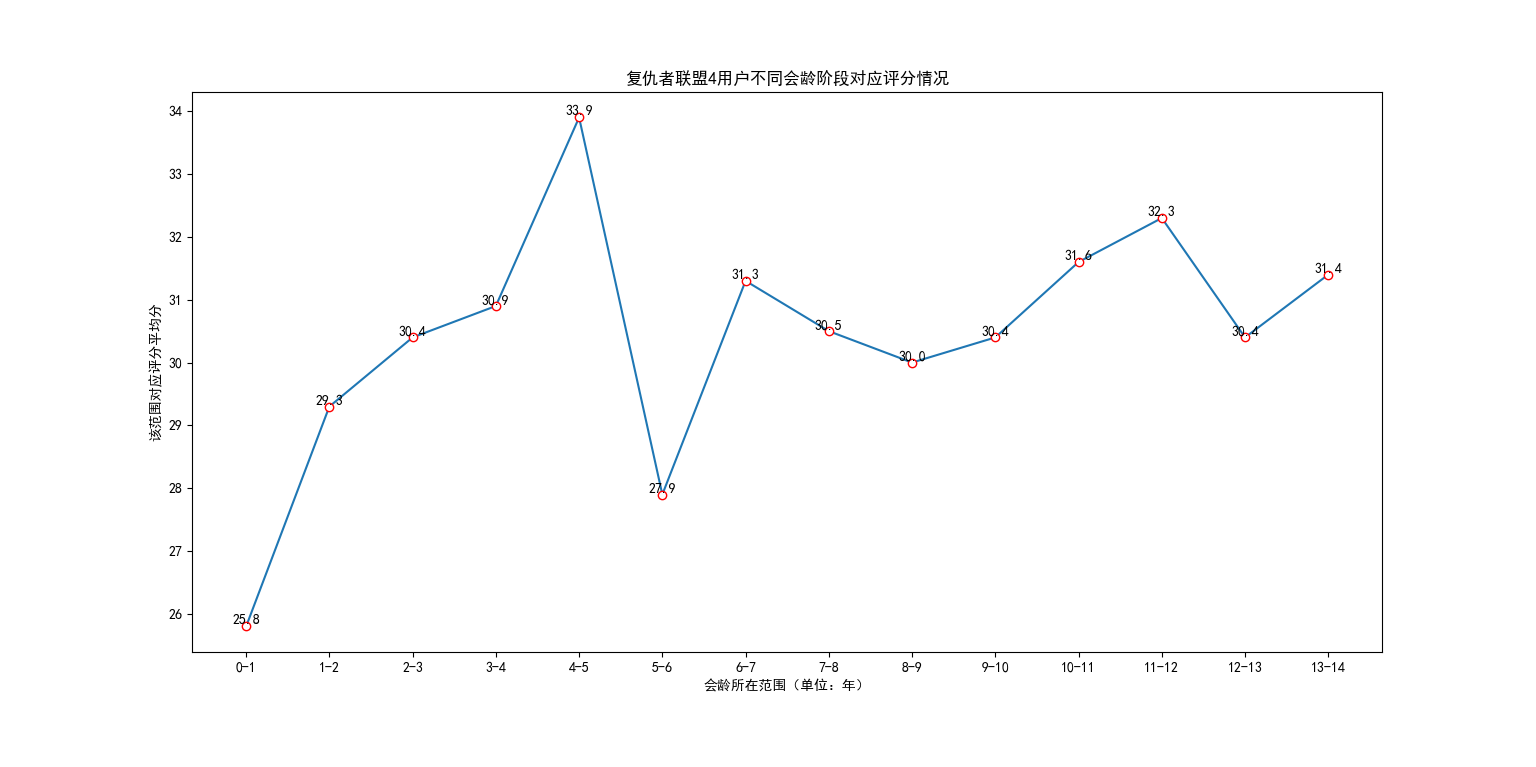
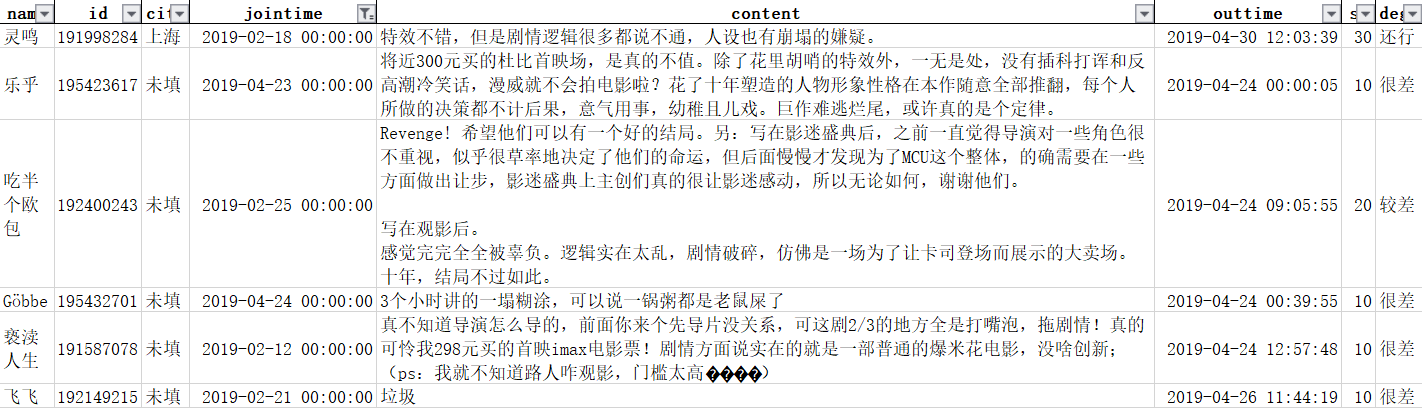
通过附件TASK13，对用户不同会龄阶段对应的评分情况进行统计，绘制出《复联4》短评用户不同会龄阶段对应评分情况折线图如图8：

图 8《复联4》短评用户不同会龄阶段对应评分情况

从图8中可以看出以下2条特征：

1. 评分最高的是会龄在4-5年之间的用户，评分最低的是会龄在0-1年之间的用户，5-6年评分也较低；
2. 其余会龄范围所对应评分情况较为相似，基本集中在29与33之间，仅相差4分。

评分最低的是会龄在0-1年之间的用户，其会龄较短，不排除有水军的嫌疑，故在附件“合并”中筛选出了join time（注册时间）为2019年的6条数据，如图9。发现此6条数据对于《复联4》都是不满意的，所打的分都比较低，但其短评皆较为用心，可以猜测此6人并非水军，且其中2019年4月才注册的用户是专门针对此影片而前来注册并发表短评的。

图 9 join time（注册时间）为2019年的6条数据

但是根据其评论可以看出来：情怀吸引用户前来观看，但同时用户对影片的要求也会更高一筹，如果掌控不好情怀的量，也是有可能招来坏口碑的。故情怀虽可以有，但更加要注重情节，注重路人是否看得懂此影片。