python 爬取 B 站“影视杂谈”系列视频并分析数据

背景

现如今，电影早已成为人们进行娱乐放松的一种选择。不同类型的电影迎合了各种人群的口味，总有一类能够被吸引。作为一名电影爱好者，初入数据分析领域，因此想选择相关内容进行数据分析。

说明

本次的数据来源源自 B 站中的“影视杂谈”类视频(https://www.bilibili.com/v/cinephile/cinecism/?spm_id_from=333.5.b_63696e6570686966c655f63696e656369736d.2),按照**视频热度排序**，爬取了 **2019-10-29 至 2019-11-05** 期间的视频相关信息。

总共爬取了 **240 页**，共 **4759 条**数据。每条数据包含（**视频标题，视频上传者(UP 主)，视频播放量，视频评论量**）

正题

A. 分析视频播放量

a) 为所有视频的播放量进行分区

视频播放量区间/10W	视频数量	区间视频播放总量	区间视频播放总量占比
4-100003	4698	16310670	48.47%
100004-200003	31	4525000	13.45%
200004-300003	10	2323000	6.90%
300004-400003	9	3117000	9.26%
400004-500003	4	1789000	5.32%
500004-600003	2	1085000	3.22%
600004-700003	1	649000	1.93%
700004-800003	1	791000	2.35%
800004-900003	1	846000	2.51%
900004-1000003	1	980000	2.91%
1200004-1300003	1	1235000	3.67%
总计	4759	33650670	100.00%

结论

由图可见，视频的播放量主要还是集中在 4-100003 区间范围内，而从 600004 到 1300003 的播放量区间，仅各有一个视频，成为佼佼者，呈现金字塔现象。这不由得让人思考，若想成为一名优秀的影视输出 up 主，是否容易成功？

很明显，这个问题从不同立场角度可以得出不同结论。

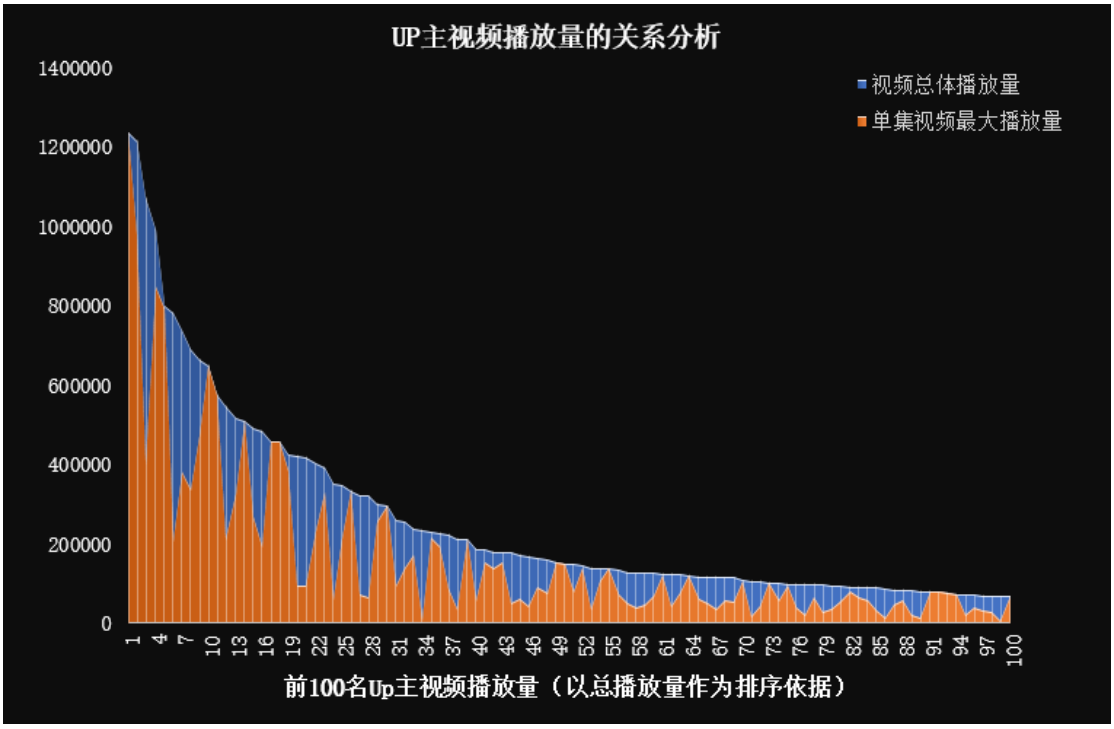
第一类，对目前已经入行且实力长期处于 4-100003 区间的范围呢，对于达到优秀的目标，需要一段十分艰难的历程也可以考虑换个领域。

第二类，对正准备入行，处于局外人的角度，这何尝不是一种挑战，也是一个机遇。当大部分人表现都不太优秀时，若是对影视充满热爱，有独特想法，完全可以进来试试，成为一匹黑马。但若长期处于不太高的播放量时，还是需要参考一下第一类建议，选择合适的。

当然仅仅从播放量上也是不能定义优秀的，不能排除有滥竽充数的现象发生，可

以参考第 D 项分析。但整体趋势还是能大体定性。

b) 对前 100 位 up 的播放量进行分析



说明

解释选取 100 位这个数据的理由：

第一， 可以参考如上所有视频播放量分区图，100 位 up 主的播放量已经涵盖了所有区间的流量数据，我们可以从中得出较为准确的结论。

第二， 对于绘图而言，数据量的减少也相对更加清晰明了。

解释加入视频总体播放量和单集视频播放量做对比的理由：

第一， 能清晰的展示每位 up 所上传的视频总体播放量及单集最大播放量的趋势。

第二， 第二，对两者的关系可以进行一个分析探讨。

结论

由图可知，总体播放量的下降是非线性的，排在前面的播放量呈阶梯性下降，越往后则趋于平缓，下降趋势减弱，可结合如上所有视频播放量分区图观察。

通过单级最大播放量的趋势也可看出，无论在任何区间内都存在有总体播放量由一集视频支撑或较为均匀分布的现象。由图上面积差异比较，两者现象产生的比例差不多为 1：1。实际计算准确得出，总体播放量靠单集播放量支撑的现象占比有 41.39%，与图上分析所得结论接近。

这一分析结果也有一定问题，因为收集的周期短，有些 up 主只产出一集视频，这种无法产生均匀分布的可能，最终结果也可能会有偏差，需要更多数据进行支撑。

这也由不得产生一个问题：究竟是需要做出一集大火视频还是追求稳定，保

证每集视频的播放量？

影视相关视频的制作想要质量有所保证，在刚开始 up 主就需要定位好自身位置。

第一类：对自身有高要求的 up 主，希望自己制作的整体视频都能有好的质量及优秀观点的输出，那么保证每集视频的播放量一定是个最好的选择，即使是刚开始没有资源的时候。追求一集大火吸引粉丝自然也可以，但无法保证路人转粉的几率。

第二类：希望快速吸引大量的粉丝的 up 主，在初期资源人力都不够的情况下，制作一集吸引眼球的视频，可能会达到效果。但长期进行之后，粉丝数量的涨势一定会趋于平缓，最终也可以考虑走第一类的路线。

B. 分析视频评论数

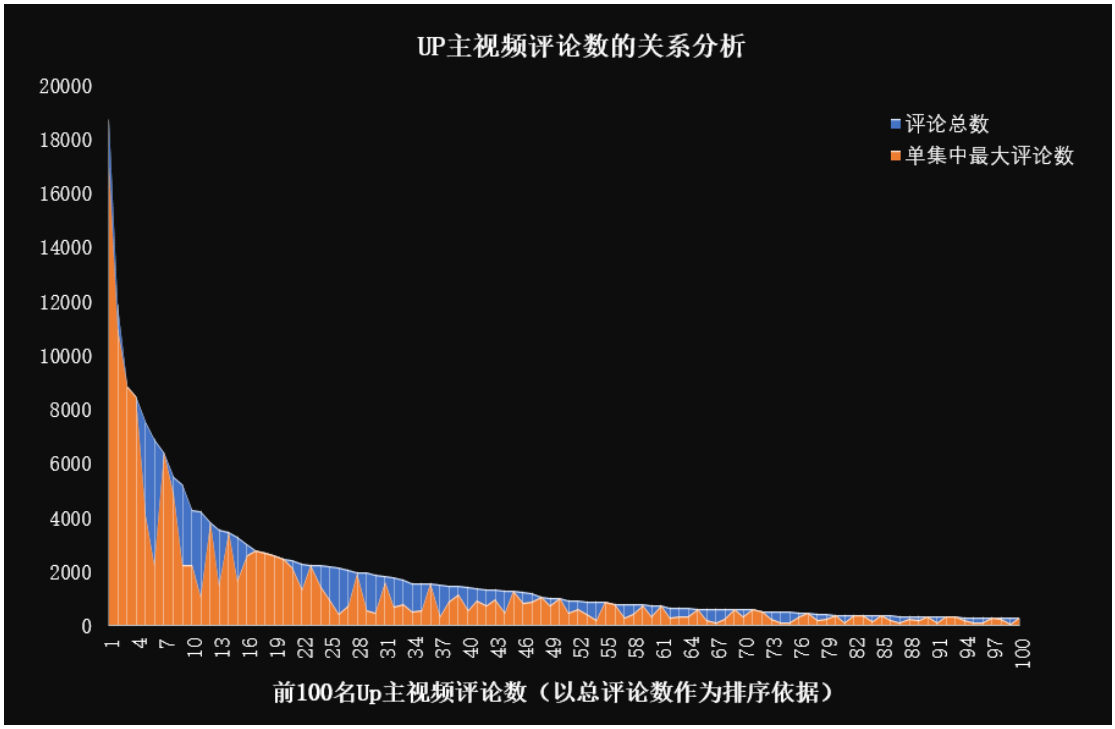
a) 为所有视频评论数进行分区

视频评论数/1000	视频数量	区间视频评论数	区间视频评论数占比
0-999	4718	96159	43.46%
1000-1999	20	28128	12.71%
2000-2999	12	29075	13.14%
3000-3999	2	7312	3.30%
4000-4999	2	8875	4.01%
6000-6999	1	6404	2.89%
8000-8999	2	17329	7.83%
11000-11999	1	11000	4.97%
16000-17000	1	17000	7.68%
总计	4759	221282	100.00%

结论

与视频流量区间图对比，在最底层的区间里视频集中数更明显，99%的视频评论数都处于 0 至 999 的区间。但结合区间视频评论数占比来看，其总计的评论数还没有占到所有的一半，而评论数前三的区间内仅仅有 4 个视频，评论数占比却接近总体的 1/4。这样的对比还是比较突出。

b) 对前 100 位 up 的评论数进行分析



说明

在此选择前 100 位评论数最高的 up 主同前面选择的前 100 位播放量最高的 up 主结果排序并不一致。

解释选择视频评论数分析的理由：视频的评论数反映了粉丝对某些电影的讨论热度以及与 up 主之间的互动热情。

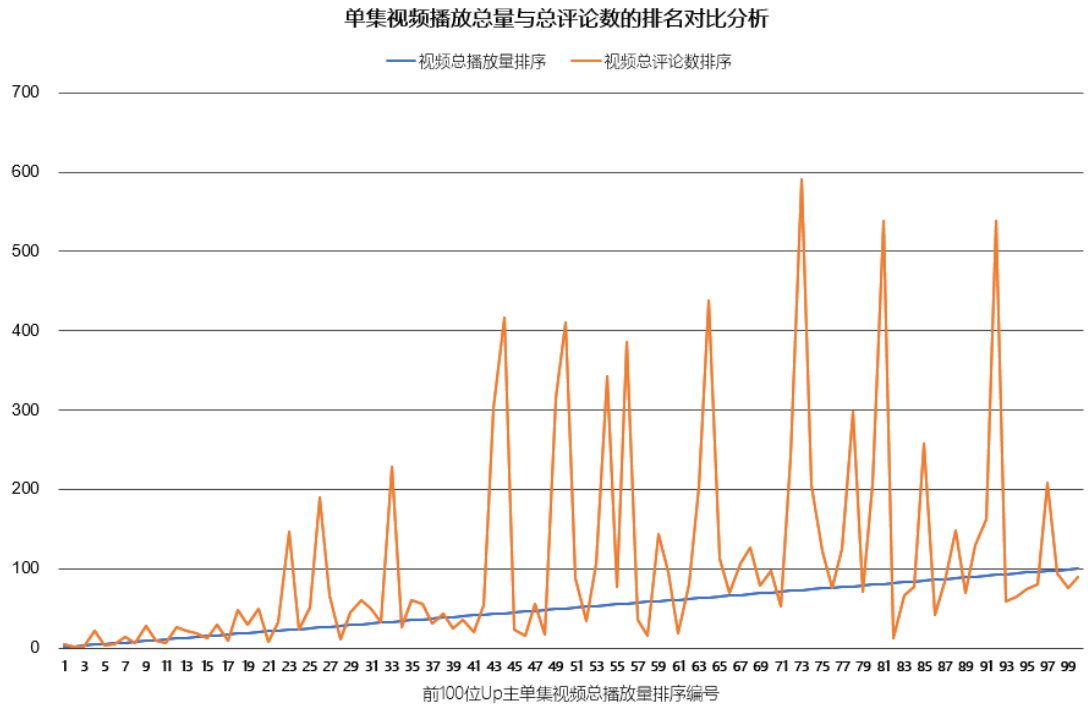
结论

虽然评论数里无法确定真实评价与灌水评价的数量，但还是能看出一些大体趋势。与播放量的下降趋势不同，总体评论数排名靠前的视频下降趋势更为陡峭，对比十分明显，后面也是趋于平稳且从 25 名左右开始，可参考如上所有视频评论数分区图。

由此也可看出粉丝对于影视类视频的反应其实很平淡，有可能因为这系列视频的类型所决定，大部分 up 主推荐的影视作品多为冷门作品或滥竽充数系列，粉丝没有去涉猎过相关内容，也无法产生强烈的共鸣，无论吐槽还是阐述观点都不太适宜。也有可能跟视频的播放量有关系，受众数量不够多，可参考第 C 项内容。

C. 结合视频播放量和视频评论数分析

a) 单集视频播放总量与视频评论数排名对比

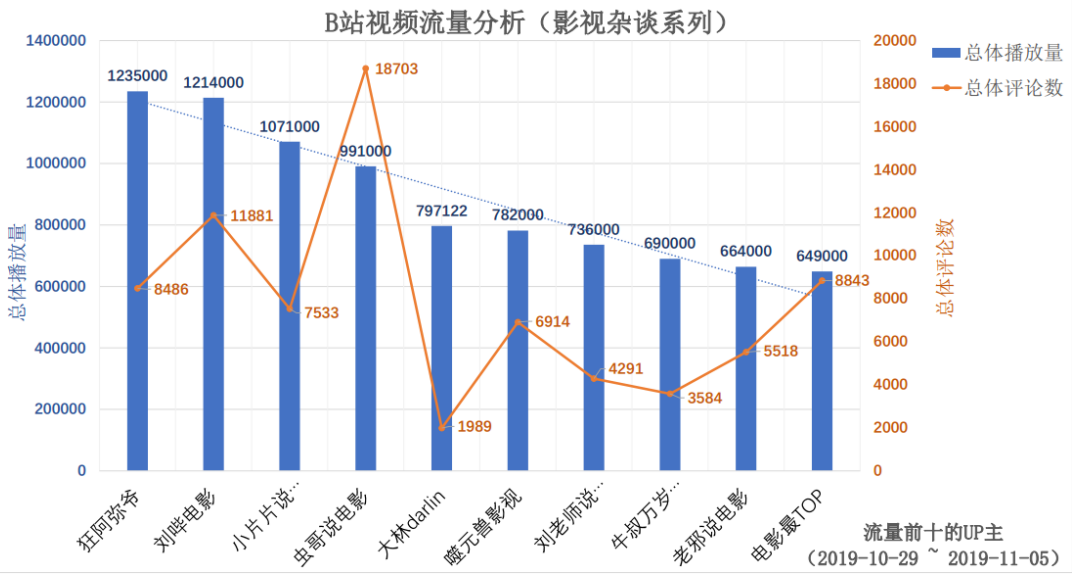


结论

在前 100 位视频总流量排序基础上，通过对比前 100 位视频总评论数的排名可以看出，视频播放量的大小与评论数的多少并不呈线性相关，在前 20 名总播放量视频里，播放量与评论数相关性较高，越往后，播放量虽然减少但其粉丝评论数却变高。

所以单看两者中的任何一个因素都不能准确分析视频质量与受粉丝喜爱程度，还是需要两者结合分析。

b) **对视频总播放量前 10 位的 up 进行分析**



说明

解释选取 10 位这个数据的理由：本次分析目的不是精确的分析出影视系列视频的某些规律及如何改进，而是一种对可能改进视频质量及增加粉丝数的探讨。

且本次收集的数据周期较短，也无法体现出明显的普适性。所以选取 10 位在本期间具有代表性的 up 主进行分析讨论。既能使图中数据更加清晰明了，也能让人对播放量和评论数有整体上的认识。

结论

由图中可看出，当视频播放量下降时，评论数并没有随之下降，排名第 4 名的 up 主有着最高的评论数，而排名第 5 名的 up 主却有着最低的评论数。

视频总播放量主要体现了视频的传播范围是否够远，受众面是否大，而视频评论总数可能体现了 up 主粉丝的黏性，互动性。通过对视频总播放量与评论数的结合分析，可以看出两者间的相关性的强度不是想象中那么高，至少在影视评论推荐类型的视频里。

当然，如果是一期热门话题电影，播放量及评论数都会出现增长，或者视频的内容质量佳且能够引发粉丝的共鸣，也有可能出现整体增长。特别是冷门系列的电影可能会更加考验 up 的讲述及分享能力，是否能完成一集优秀的视频取决于 up 主的前期积累（粉丝积累，素材积累，创意积累等）。

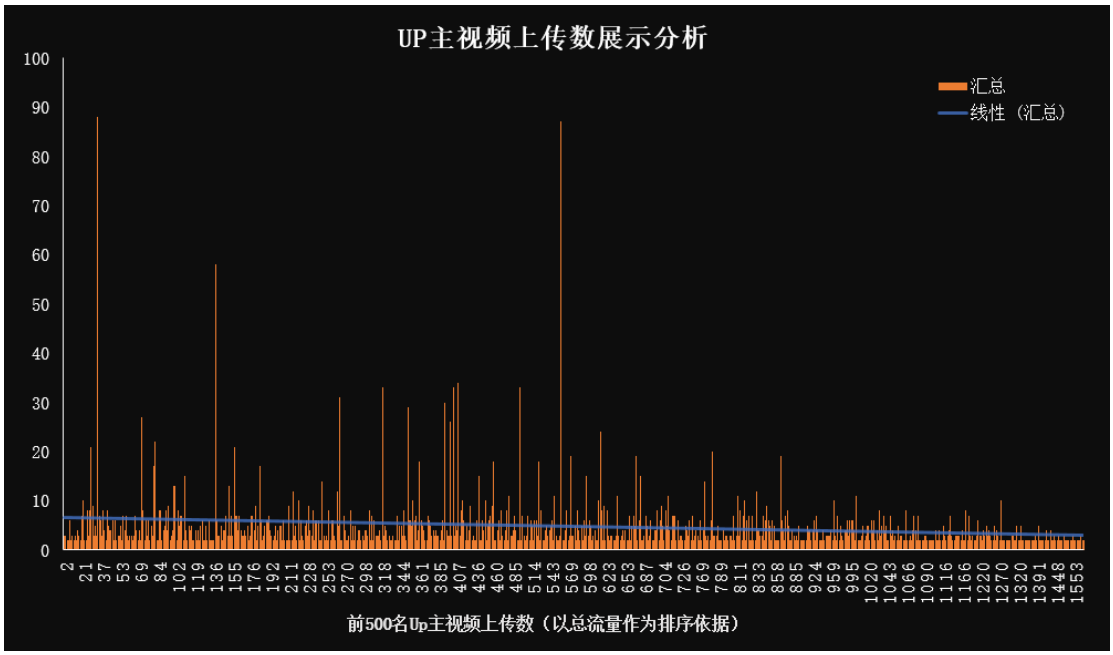
这也可以引出一个问题：衡量一个视频是否达到预期，是否较为成功的标准是什么？

这个问题很明显是因人而异，每位 up 可能追求的不一样，希望收获到的东西也不一样。但大部分都还是会希望能够受更多人的喜爱与接收。

在这一点上，播放量和评论数可以作为较为重要的评判依据。播放量的成功可能不代表着视频质量一定就是好的，或许是噱头成分较多，评论数也会灌水。但一个较为优秀的视频且能被大众看见的时候，播放量达到一定程度其评论数也一定能有所增长且高于平均水平。这或许是一条能够作为判断的依据，万事不唯一，但一些基本规律还是可以参考。

如果目前还处于初期的 up 主，可能即使视频优秀也无法被传播开去，自然这样的评判标准也失去了意义。而当前，能够参考的依据则可以从粉丝数量的增长率及与自身过往视频的播放量进行对比分析，这样一定能找到部分答案。

D. 分析视频上传数



说明

解释选取前 500 位这个数据的理由：排名前 500 的视频上传数基本覆盖了所有 up 主能生产的值，最低即是 1，本次目的主要是查看是否有量产极高的现象发生并分析。

结论

由图，以线性趋势看，整体 up 主的视频上传数还是较为稳定，基本维持在 1 至 10 的范围内。随着播放量的减少，整体也呈现了略微下降趋势。这也能间接反映出，播放量长期上不去，确实对 up 主制作视频有所影响，是 up 主持续创作的动力之一。

也确实存在过于显眼的异类，最高的上传数达到了 88 集视频，up 主的更新频率过高。在这较短的收集期间里，这样的数据不难让人觉得有浑水摸鱼的情况，试图依靠高产来提高播放量，达到快速吸引人群的目的。因此，这一类极端情况在对视频总播放量分析的时候，暂时还未排除，可以能会导致一定的偏差。

高产虽然不意味着一定没有质量，但过于高产对质量一定是有所影响的，或者考虑是否有抄袭他人视频的嫌疑。这样的 up 主若只顾眼前的利益，自然是走不长远，即使播放量上去了，其他的评判指标（评论数，粉丝数等）也照样能够反映问题。只为做博人眼球的视频，而不顾内容的投机行为，能带来的收益是微乎其微的。

思考及问题

在最后，想再对自己本次分析做个总结。考虑一些在本次分析中并未收集或考虑周围的问题。

在未收集的数据中有些数据我认为值得加入进行对比结合，产生更准确的结果。例如以下几点：

1. 每位 up 主的粉丝数量可以与评论数结合，进而分析出 up 主与粉丝间的互动是否良好

2. 视频创建的时间。虽然本次收集周期短，但每个视频的上传时间还是有所不同，其播放量和评论数也还是有所差异，对结果有所影响
3. 视频当天发布时间点。由此可以分析出 up 通常会处于当日哪个时间段，进而分析出不同时间段，对播放量可能造成的影响。
4. 视频中涉及的电影相关信息，类型，上映时间，评分等等。通过分析电影信息可间接分析出 up 主是否有紧跟热点，或更偏向追求经典，与其视频的播放量，粉丝数，评论数的关系。

本次分析还有很多可以完善的细节与内容。得出的规律或建议也仅限于此次收集的数据，若想要真正认识到数据的力量，还需要更多的练习与挑战，收集更大范围内的数据，考虑更多的因素，发现其中隐藏的规律与事实，值得用更多时间去探寻挖掘。