数据科学导论 项目报告

选择数据：`bilibili_data.csv`

组员	学号
李乐平	12112627
梁瑞玮	12111219
何其佳	12111211

B站up主抽样数据分析：up主行为研究概述

导言

随着科技的发展与社会的进步，在近几年中，哔哩哔哩弹幕网平台（以下简称B站）逐渐走向成熟，成为了中国年轻世代高度聚集的文化社区和视频网站。随着B站用户的增多和up主（B站up主的称呼）数量的上升，B站一定程度上反映了中国年轻一代的价值取向，从而对B站大数据研究的重要性也就愈加突显。通过对B站up主行为研究分析，我们能够更加清晰地反映当前年轻人的心理需求以及兴趣偏向，up主的部分行为和粉丝数量关系，以此来更加清晰地洞察当前主流价值取向并对新老up主制作视频、吸引粉丝提供相应的建议。

目录

01数据概要、分类与处理思路

1.1 数据概要

在提供的数据中（将该数据集称为`old_bilibili`），共收集了8360位up主的个人信息，单个up主信息如下：

mid	follower	sex	master	album
个人编号	粉丝数	性别	代表作个数	发布相册数
article	**channel**	**time_ave20**	**play_ave20**	**weibo**
发布文章数	发布频道数	近期视频平均时长	近期视频平均播放量	是否关联微博
wx	**qq**	**taobao**	**mail**	**tiktok**
是否关联微信	是否关联QQ	是否关联淘宝	是否关联邮箱	是否关联抖音

mid	follower	sex	master	album
redbook	**self_tags**	**video_tag_combine**	**video_max_ratio**	**video**
是否关联小红书	自我标签	视频分区	主类视频比例	作者视频数量

该数据集获取时间为2019年，距今已有一定时间。为了进行纵向对比研究，我们通过谷歌的web Scraper功能获得了以上8360位up主在2022年11月的相关数据（数据集称为`new_bilibili`）：

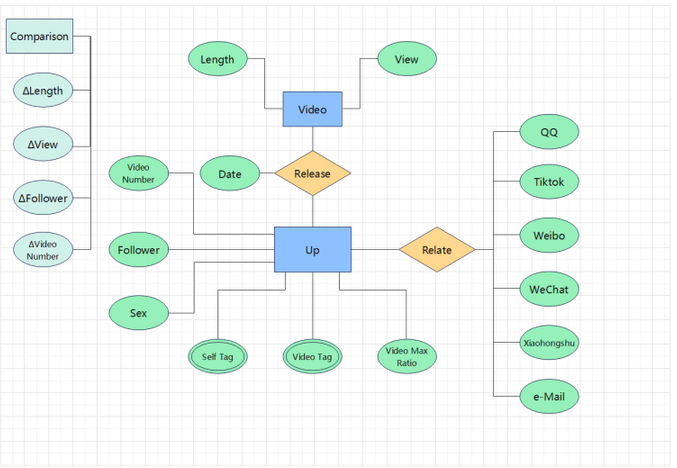
name	fans	video_numbers	recentVideo_publishedTime
Up主名称	当前粉丝数	视频的总数	最新视频发布时间
date	**length**	**id_**	**play**
近期视频的发布时间	近期视频的时长	个人编号	近期视频播放量

1.2 数据分类与处理思路

在数据处理上，我们小组主要围绕**粉丝数**这一因变量，探究各个自变量对其的影响。

我们先把数据分为两大主类，**过去的**和**现在的**，对于过去的变量先采取在同一时间节点横向展开的方式处理数据，探究`old_bilibili`中各变量对粉丝数的影响；然后对数据纵向展开，以前后时间节点上的粉丝数变化为因变量，探究up主的活跃度、视频发布数量等自变量对其的影响。

综合利用以上数据，我们小组的数据处理思路如下图所示



我们的数据分析工作可总结为以下两点：

1. 对`old_bilibili`中up主的基本信息进行统计分析
2. 对比`old_bilibili`与`new_bilibili`中的相关数据并进行分析

02 数据处理

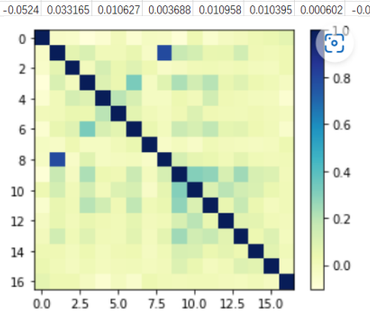
2.1 不同变量之间的关系

2.1.1 基本信息数据间相关系数分析：

首先我们对`old_bilibili`中各项数据的相关系数进行了研究，由相关系数矩阵和热量图可知：

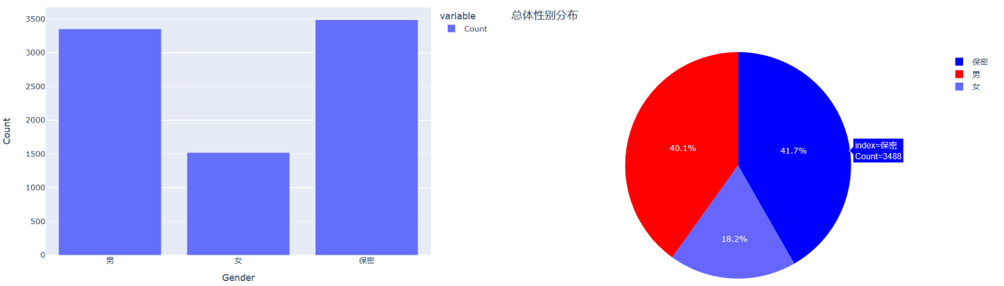
`play_ave20`和`follower`（0.786），`wx`和`weibo`（0.296），`qq`和`weibo`（0.278），`weibo`和`master`（0.222），`qq`和`master`（0.202），`mail`和`weibo`（0.246），`article`和`album`（0.197），`taobao`和`wx`（0.195），`mail`和`qq`（0.173）的关联度较高（≥0.15）。其余的量相关系数的绝对值较小，说明关联度不大。

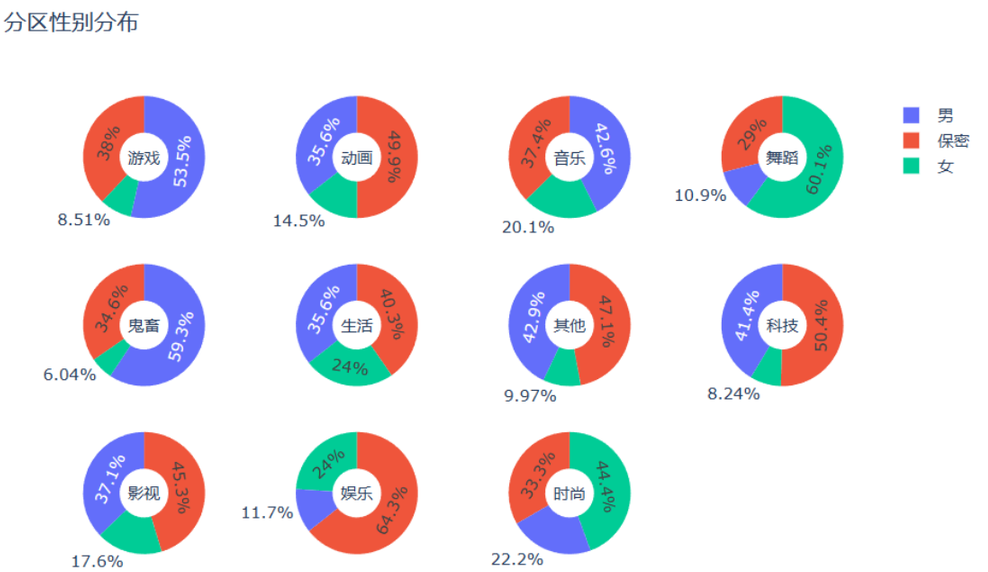
mid	follower	video	master	album	article	channel	time_ave2	play_ave2	weibo	wx	qq	taobao	mail	tiktok	redbook	video_max_ratio	index
1	-0.04183	-0.05082	-0.10263	-0.07136	0.004977	-0.07067	-0.02318	-0.03108	-0.09108	0.029736	-0.06358	-0.01732	-0.0021	0.025157	0.043585	0.068717783	mid
-0.04183	1	0.066581	0.105413	0.006819	0.007061	0.049929	-0.03588	0.785638	0.167802	0.129093	0.051706	0.052689	0.137164	0.019058	0.01321	0.029027607	follower
-0.05082	0.066581	1	-0.00596	0.123503	0.046169	0.084099	0.027373	-0.04043	-0.02058	0.012499	0.017886	0.001479	0.01527	-0.00538	-0.01379	0.028506498	video
-0.10263	0.105413	-0.00596	1	0.094419	0.039828	0.321064	-0.02325	0.078409	0.221738	0.134232	0.202017	0.052099	0.085241	0.03347	0.014434	-0.019545533	master
-0.07136	0.006819	0.123503	0.094419	1	0.196616	0.115833	-0.02132	-0.02117	0.032017	0.002792	0.135225	0.022144	0.011385	-0.00509	-0.01092	-0.068703394	album
0.004977	0.007061	0.046169	0.039828	0.196616	1	0.071365	0.009043	-0.01389	0.019314	0.098415	0.017996	0.010377	0.009191	0.019017	0.002589	-0.013790128	article
-0.07067	0.049929	0.084099	0.321064	0.115833	0.071365	1	0.044981	-0.00518	0.159519	0.120346	0.184354	0.062296	0.078204	0.02904	0.007638	-0.052401928	channel
-0.02318	-0.03588	0.027373	-0.02325	-0.02132	0.009043	0.044981	1	-0.041	-0.01047	-0.04099	0.014034	-0.00257	-0.01029	-0.01718	-0.00882	0.033165256	time_ave20
-0.03108	0.785638	-0.04043	0.078409	-0.02117	-0.01389	-0.00518	-0.041	1	0.118658	0.06082	0.044586	0.011333	0.066975	0.015043	0.002695	0.01062708	play_ave20
-0.09108	0.167802	-0.02058	0.221738	0.032017	0.019314	0.159519	-0.01047	0.118658	1	0.298513	0.278034	0.113206	0.246289	0.113681	0.097481	0.003687859	weibo
0.029736	0.129093	0.012499	0.134232	0.002792	0.098415	0.120346	-0.04099	0.06082	0.298513	1	0.103292	0.194897	0.136022	0.104176	0.045916	0.010958333	wx
-0.06358	0.051706	0.017886	0.202017	0.135225	0.017996	0.184354	0.014034	0.044586	0.278034	0.103292	1	0.050106	0.172691	0.040351	-0.00999	0.010394771	qq
-0.01732	0.052689	0.001479	0.052099	0.022144	0.010377	0.062296	-0.00257	0.011333	0.113206	0.194897	0.050106	1	0.074293	0.067155	0.011208	0.000601625	taobao
-0.0021	0.137164	0.01527	0.085241	0.011385	0.009191	0.078204	-0.01029	0.066975	0.246289	0.136022	0.172691	0.074293	1	0.032593	0.092866	-0.00906297	mail
0.025157	0.019058	-0.00538	0.03347	-0.00509	0.019017	0.02904	-0.01718	0.015043	0.113681	0.104176	0.040351	0.067155	0.032593	1	0.076695	0.007675273	tiktok
0.043585	0.01321	-0.01379	0.014434	-0.01092	0.002589	0.007638	-0.00882	0.002695	0.097481	0.045916	-0.00999	0.011208	0.092866	0.076695	1	-0.018282959	redbook
0.068718	0.029028	0.028506	-0.01955	-0.0687	-0.01379	-0.0524	0.033165	0.010627	0.003688	0.010958	0.010395	0.000602	-0.00906	0.007675	-0.01828	1	video_max_ratio



2.1.2 不同分区的性别比例

将`old_bilibili`的`sex`信息进行统计分类，将其分别绘制成直方图与饼状图，再对主题进行分类，对各分区的up主性别成分进行分析，具体图形如下：





由图1、2数据可得，在公开的数据资料中，男性的占比达到了68.2%，女性为31.8%，约为女性数量的2.14倍。但在整体数据中，41.2%的up主都不愿意公开自己的性别资料。

由性别分区分布可得，男女数量比差异在**鬼畜**、**游戏**、**科技**等领域达到最大，在**舞蹈**、**时尚**、**生活**领域达到最小，其中舞蹈和时尚两个领域的女性比例高于男性。一定程度上反映了男性女性的整体的兴趣集中点的方向。

2.1.3 标签词频统计

提取`new_bilibili`中的`self_tags`类，将全体数据合并程一个`txt`文档，共计8889个个人标签词汇，经微词云处理后，生成的词云图如下：

因为个人标签的自由度较高，对于低频词汇的统计的价值和意义不大，其中排名前20的个人标签词汇和对应数量如下：

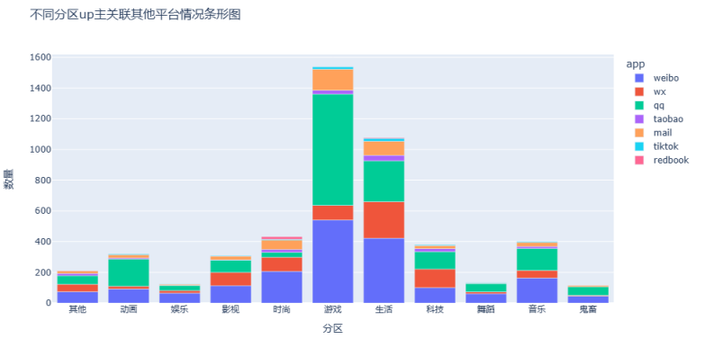
游戏	解说	搞笑	美食	电影	美妆	视频	鬼畜	音乐	动漫
240	112	104	95	94	81	72	62	61	65
生活	**时尚**	**主播**	**字母组**	**实况**	**原创**	**科技**	**翻唱**	**日本**	**MMD**
52	48	45	45	44	44	43	42	42	40

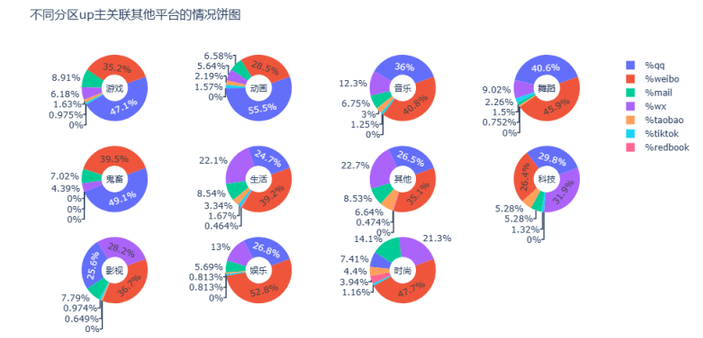
由词频分析可得，当下年轻人的主要兴趣集中在**游戏**（游戏、解说）、**生活**（生活、时尚、美食、美妆）、**音乐**（音乐、翻唱）、**动漫**（动漫、MMD、日本）上。

2.1.4 up主关联号分析

1. 研究不同分区关联其他平台人数与占比

我们对不同分区的有关联其他app平台的up主进行了研究，通过绘制条形图和饼图考察每个分区中不同关联app所占的数量和比例。（备注：weibo：微博；wx：微信；qq：QQ；taobao：淘宝；mail：邮箱；tiktok：抖音；redbook：小红书）



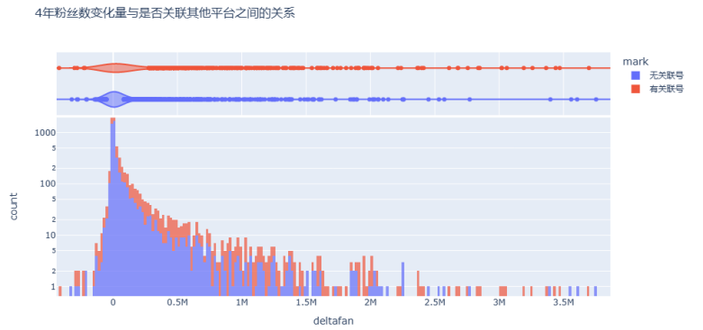


分析：

1. 在游戏区、动画区、鬼畜区中，可见关联QQ的up主所占比例基本达到50%，归因为QQ作为青少年聚集的社区，游戏、动画、鬼畜等与ACG（Animation, Comic & Game）相关话题更容易在QQ中得到关注；
2. 音乐区、舞蹈区、生活区、影视区、娱乐区中，可见关联微博的up主所占比例较大，归因为微博作为普及度较高的公民媒体，音乐、舞蹈、生活、影视、娱乐等偏大众性的话题更容易在微博得到传播；
3. 关联微博的up主和关联QQ的up主在各个区中占比都比较大，可见bilibili平台与QQ、微博平台关联性、互动性较高；
4. 关联微信的up主大多数出现在生活区、科技区、影视区、时尚区，而较少出现在游戏区、动画区、鬼畜区等，归因为微信用户年龄分布偏大，适合大众性话题的传播，不适合ACG等话题传播；
5. 在所有区中只有极少数up主关联了抖音、小红书、淘宝，说明此类app与bilibili平台关联度并不高。

2. 关联其他平台是否有助于粉丝量的增长

我们研究并绘制了四年up内主增粉量与up是否关联了其他平台之间的关系：



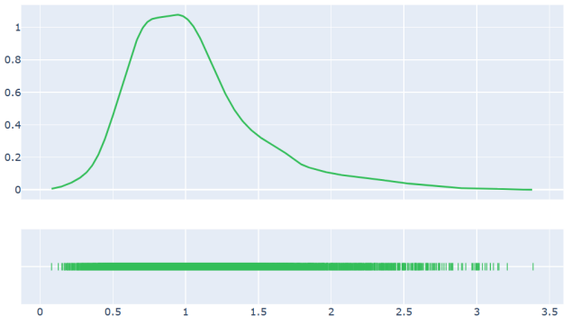
可见**有关联号的up主在四年内增粉数较大，增长潜力也更明显。**归因：**关联其他平台可在一定程度上提高up主的知名度，使其在bilibili上获得更多的关注。**

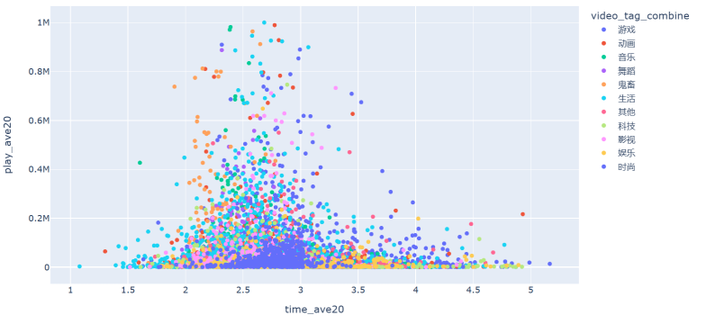
2.1.5 视频长度和视频播放量之间的关系

我们研究了up主的**视频长度分布**以及up主的**视频长度和视频播放量之间的关系。**将 `video_ave20` 取对数（log10）后绘制图形如下：

如下图所示，up主的视频长度呈现先增后减的趋势，大部分的up主的视频时间集中在10分钟上下（根据处理公式换算得到，10^t）。

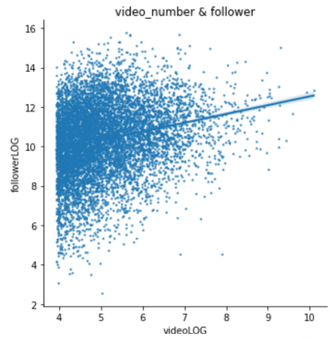
如右图所示，在清理掉播放量大于一百万的视频后（极少数据，以突出散点图大部分数据特征），**$sup(视频播放量)$和$log_{10}(视频时长)$**基本呈现先升后降的趋势，表明大多数视频播放量大的视频集中在中间，**其时长大约在1分40秒-16分40秒之间**（图中横坐标2-3区间）。**反映了人们对视频时长的偏赖程度。**但**同时在上述（2-3）区间内仍有大量视频的播放量较低，与左图信息结合，表现出这个区间的视频的竞争压力也相对较大。**

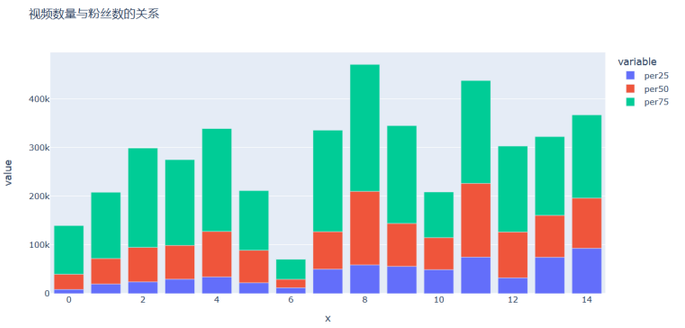


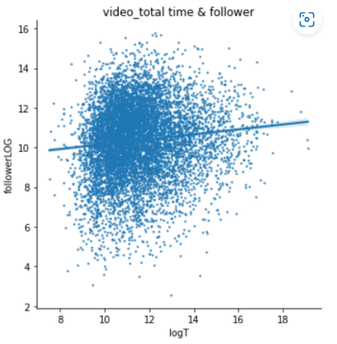


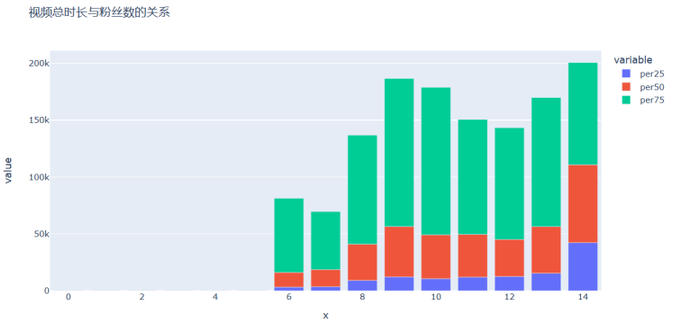
2.1.6 up主视频与粉丝数的关系

up主视频的数量与粉丝数具有一定的关系。我们分别对up主视频的数量，up主视频的总时长进行了统计，如下图所示。左图为视频数量、总时长与粉丝数关系的散点图（数据均取了对数），右图为条形图，两者均呈现波动上升趋势。由此得，在视频质量相同的情况下（基数足够大后将视频质量视为相同），**视频发布越多越有利于粉丝数的增长。**





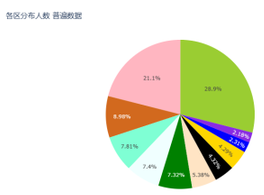


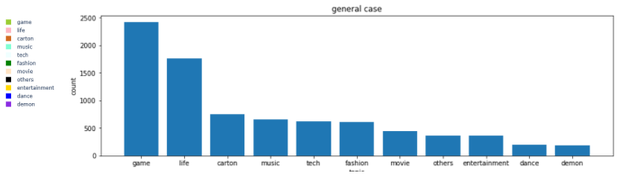


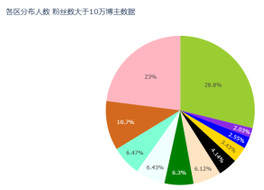
2.1.7 up主在不同分区的分布情况

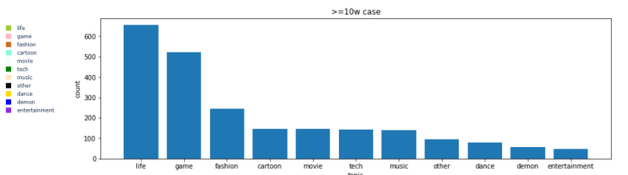
B站一共分为 游戏 、 卡通 、 鬼畜 、 音乐 、 舞蹈 、 生活 、 科技 、 电影 、 娱乐 、 时尚 、 其它 十一个分区，每个up主都会根据其特点被归入到一个区中，受到B站文化倾向的影响，up主在各个分区的数量具有明显差异，通过统计up主的的分区情况，能够得出一些基本的青年价值兴趣取向：

在受到兴趣和他人的影响，很多人开始运营自己的up主账号，但视频发布往往具有低质、随意的特点。在讨论up主分区的情况时，我们将up主做了两类统计：高质量up主（粉丝数>=10万）和全部up主（不对数据做任何筛选）。up主在各区所占比例图如下：









大部分的up主主要集中在 生活 、 游戏 两大主要分区，对比前后两组数据，对各个主题的排序的升降对比，得出各主题在up主精英化后排名的变化情况：

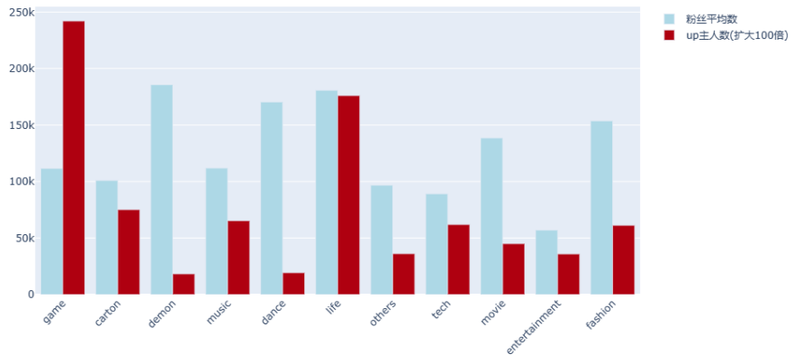
主题	时尚	电影	生活	舞蹈	鬼畜	其它	科技	卡通	游戏	娱乐	音乐
Δ排名	+3	+2	+1	+1	+1	0	-1	-1	-1	-2	-3

通过排名的对比变化，我们发现 **时尚、电影**等数据的排名具有一定上升，而**游戏、娱乐**等排名具有一定下降，说明想要做好**时尚、电影**等偏大众性的文化话题具有一定的难度。但整体的排名和类型比例没有大幅变化的情况，表明在各个类型中的up主发布视频的平均质量和水平差别相对较小。

因为每个分区的视频制作区的up主数量的差异明显，就会有不同分区的**内卷程度**不一的现象，我们将平均粉丝数来判别每个分区的内卷程度，平均粉丝数越多，说明内卷程度越低，我们能够绘制出以下图像

$$-\text{内卷程度} \propto \text{潜在粉丝数量}/\text{潜在竞争对手} = \text{分区的全部博主全部粉丝}/\text{该区全部博主} = \text{平均粉丝数}$$

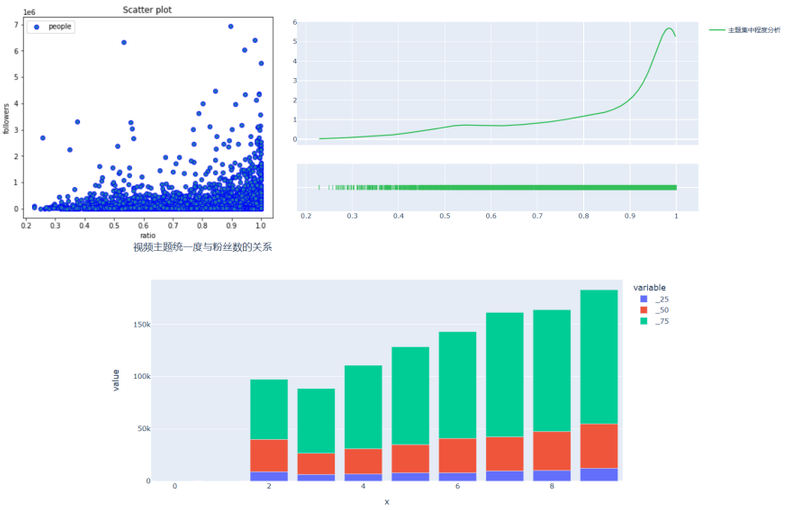
我们惊人地发现，位于平均数（127 k）以下的和以上的主题与上方排名变化为正和负的主题保持着完全一致性，**游戏、卡通、娱乐**等内卷程度严重，**舞蹈、时尚、生活**等内卷程度较低，说明在2019年，**舞蹈、时尚、生活**区的发展空间比较大。



2.1.8 视频主题集中程度与粉丝数的关系

对于up主视频制作，有的up主视频制作相对集中，有的up主视频制作相对杂乱分散，哪种情形更有利于吸引粉丝？我们对这一数据展开了分析：

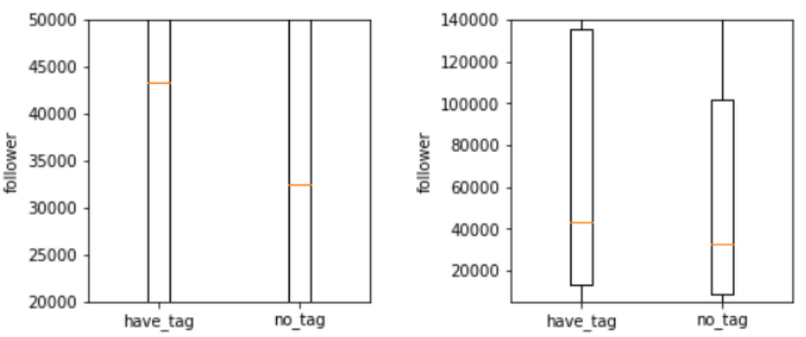
首先做出散点图，得到数据的分布情况；再绘制各个区间的分布情况：将整个数据（0-1）分成10个区间，并将每个区间的粉丝数取上、中、下位数，绘制直方图，最终得到的结果如下图所示：



由散点图和曲线图可得，大部分的up主视频较为集中，56%的up主的主题集中度都达到了90%以上，而主题集中度小于20%的up主的数量为0；由条形图，除集中度为20%-30%阶段的数据外，25%、50%、75%处的粉丝量都随着集中度的增加呈现增长态势（集中度在20%-30%的数据量只有25个，不稳定性较大，可不予考虑）。因此，**一定程度上提升up主的主题集中度有利于粉丝数的增加。**

2.1.9 贴标签对up主粉丝数的影响

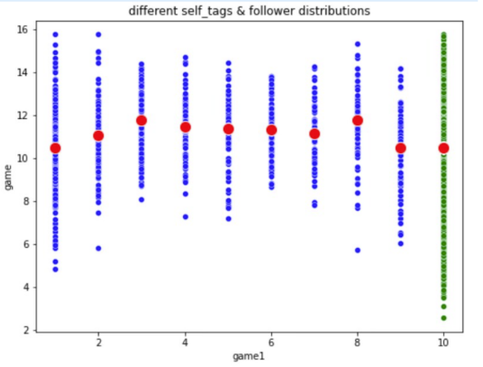
衔接2.1.2, 我们完善对标签信息的分析，在词频分析后，进一步探究其与粉丝数的关系。贴标签是否与粉丝数具有一定关联？我们首先对up主是否具有`self_tags`进行分析，放大后的`boxplot`如下所示：



由数据得，在贴了标签的情况下，up主的`boxplot`的四分位数对应的指标均高于不贴标签的up主，其中中位数之差达到了10841.5个粉丝量，整体可得出结论：贴了标签的up主平均粉丝数将会高于不贴标签的up主，推测原因为贴了标签更有利于用户精确定位up主特征和方向，同时贴了标签的up主在行为上往往更加注重细节与认真。

于此同时，对于前期调研的高频词汇，我们截取了前九项作数据分析，探究高频词汇（即用户喜欢看的类型）在标签中的出现是否和粉丝数呈现正比关系，绘制表格如下：蓝点为各个数据（类型-粉丝数（取对数））的散点图，红点为各类型中位数取值，**最后一列为全体数据集，是对照组**，比对发现，除游戏外，全体数据的中位数均比对照组略高，且低粉丝数的情况在贴热门标签后出现的次数明显低于对照组。up主的水平整体呈现层次不齐状态，贴标签一定程度上对粉丝数的增长略有积极的影响。

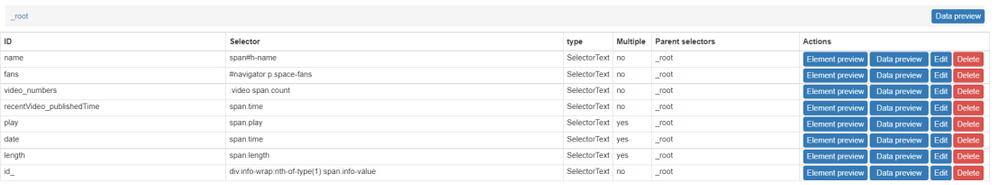
	game	explain	fun	food	movie	make_up	video	demon	music	whole
0	10.460935	11.052524	11.75453	11.440169	11.389118	11.321511	11.168137	11.78287	10.489912	10.482051



2.2 时间上的数据的对比

2.2.1 新数据获取

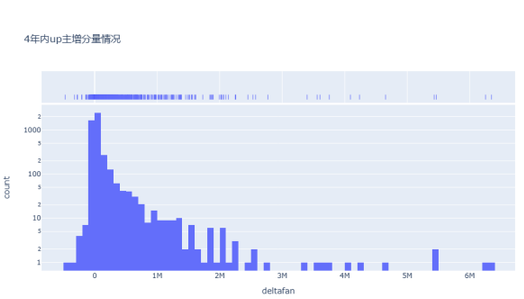
我们采用谷歌浏览器的插件Web Scraper，对获取所需要的数据。在Web Scraper 内显示的数据如下。得到`new_bilibili`的数据，因获取的up主均为`old_bilibili`中的up主，只是时间上的不同，故能够对前后的数据实现对比分析。

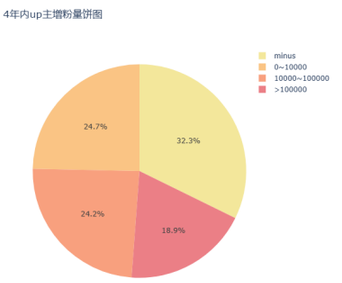


_root　　Data preview

ID	Selector	type	Multiple	Parent selectors	Actions
name	span#h-name	SelectorText	no	_root	Element preview / Data preview / Edit / Delete
fans	#navigator p.space-fans	SelectorText	no	_root	Element preview / Data preview / Edit / Delete
video_numbers	.video span.count	SelectorText	no	_root	Element preview / Data preview / Edit / Delete
recentVideo_publishedTime	span.time	SelectorText	no	_root	Element preview / Data preview / Edit / Delete
play	span.play	SelectorText	yes	_root	Element preview / Data preview / Edit / Delete
date	span.time	SelectorText	yes	_root	Element preview / Data preview / Edit / Delete
length	span.length	SelectorText	yes	_root	Element preview / Data preview / Edit / Delete
id_	div.info-wrap:nth-of-type(1) span.info-value	SelectorText	no	_root	Element preview / Data preview / Edit / Delete

2.2.2 up主增粉量简析

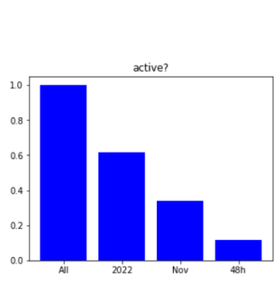


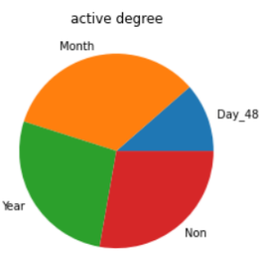


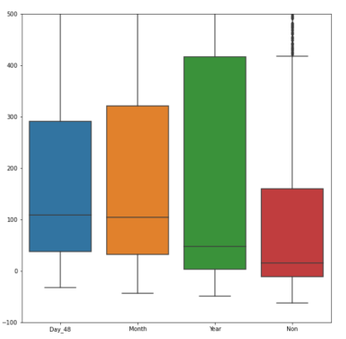
由图可得，**数据集中有32.3%的up主的粉丝量出现了负增长（minus），也有约40%的up主粉丝数增长明显。**

2.2.3 up主活跃度变化研究

基于`new_bilibili`，我们以up主的信息分为了以下四类：全体调查up主、在2022年仍发布视频的up主、在11月份仍发布了视频的up主和在48小时内发布了视频的up主（四个类型为包含关系）。由数据得出，**28%的up主已在22年未发表任何视频；34%的up主为轻度活跃（月发表），11%的up主仍然极度活跃，在48小时内由发表视频的痕迹。**







对up主的活跃度与粉丝数的分析，我们将四类数据做取补（如：`boxpolt`“month“一栏只包含在11月发表且在获取数据时48小时内未发表视频的up主）后做出`boxplot`，纵坐标为数据粉丝数增长的百分比：

$$\text{当前粉丝数量}/\text{过去粉丝数量} * 100$$

如上图所示，”完全不活跃“和”年更“的up主的粉丝增量明显低于”月更“和”日更“的。”月更“和”日更“的up主的增分量差距相对较小，”月更“up主在75%分位数的值甚至高于”日更“up主。推测**”日更“的up主更加注重视频的频率与数量而”月更“的up主更加注重视频的质量。两者均有利于粉丝数的增长，故最终的数值相差不大**。

2.2.4 各分区发展潜力分析

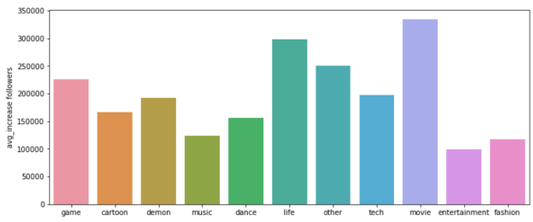
依照up主分区将up主进行分类，然后再对up主进行粉丝数增长调研研究，对比粉丝数增长的差异，我们能够得出不同领域的发展潜力的大小。

$$\text{潜在发展潜力} \propto \text{粉丝增长总数}/\text{博主人数}$$

受到活跃度的影响，活跃度较高的up主的增粉量的代表性比活跃度较低的up主的增分量对当前情况预测的代表性更强，因此我们对时间轴上各数据进行权重的分配，优化发展潜力的评估方式：

$$\text{潜在发展潜力} \propto (\text{日更博主增粉} * 0.4/\text{日更博主人数}) + (\text{月更博主增粉} * 0.3/\text{月更博主人数}) + (\text{年更博主增粉} * 0.2/\text{年更博主人数}) + (\text{不活跃博主增粉} * 0.1/\text{不活跃博主人数})$$

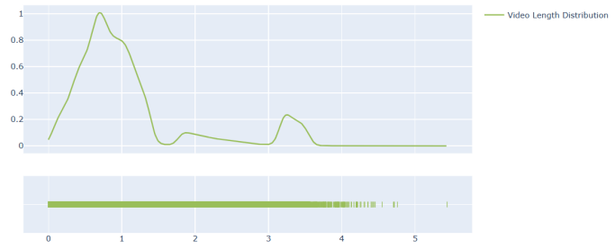
绘制出的图像如下：



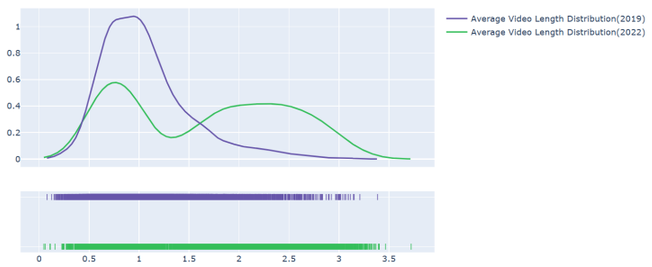
由此数据得出，**当前电影、生活、其它、游戏分区的发展潜力较大，而娱乐、时尚、音乐、舞蹈的发展潜力较小。**

2.2.5 up主的视频时长分析

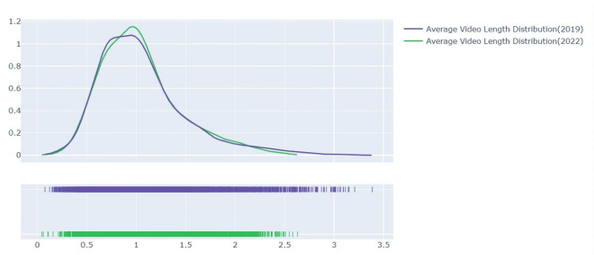
在*new_bilibili*中，我们统计了2022年up主视频时长的分布，如下图中左上图所示。可看出视频时长存在两个峰值，第一个峰值位于约5分钟处，而第二个峰值位于约2000分钟处，是获取数据时直接获取了一整个合集的视频总共时长导致（合集功能于2021年上线，可以收录一位up主一系列的视频），故不对第二个峰值进行研究。可见大多数视频时长都集中在5分钟左右。



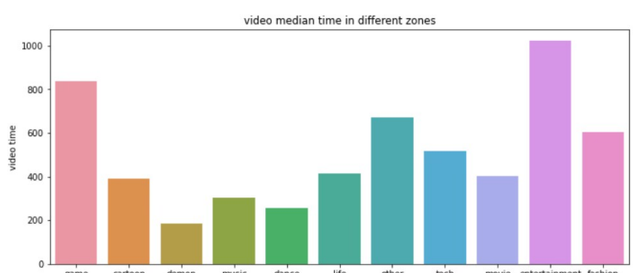
我们进而研究对比了2019年和2022年的视频平均时长，发现由于受到合集的影响，2019年视频时长和2022年视频时长在30mins后不具有较高的可比性，但从图中可大致看出合集在*new_bilibili*长视频中的比例较高，对后面的分析进行指导。



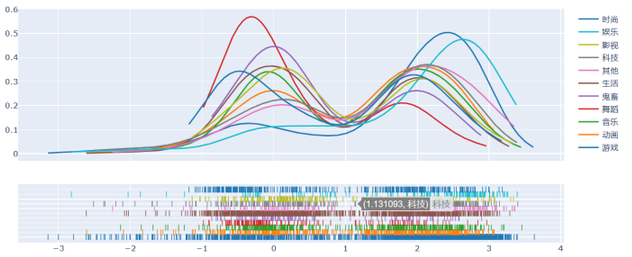
剔除掉可能被合集影响的数据，我们又比较了2019年和2022年视频长度的差异，并绘制了下图中右上图。可知2019年视频长度与2022年视频长度并无太大差异，**表明在短视频盛行的时代，B站长视频生态仍可缓冲短视频潮流的冲击。**



我们进而研究了2019年不同分区的视频时长分布情况，可见**游戏区、娱乐区、时尚区的平均视频时长较长，而鬼畜区、音乐区、舞蹈区等视频平均时长较短。**



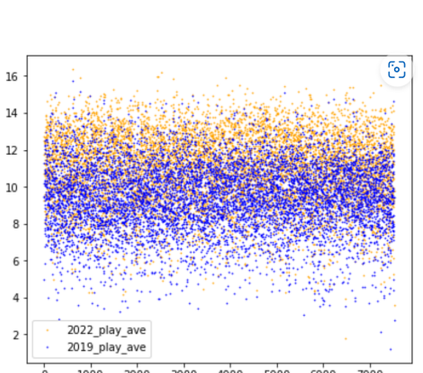
最后，我们比较了2019年、2022年视频时长的变化，下图横轴表示2022年各区平均时长-2019年各区平均时长，曲线为其密度分布曲线。观察第一个峰值，可见**舞蹈区、时尚区、游戏区的视频时长有所缩减，影视区和科技区的视频时长有所增长**，其他基本不变。观察第二个峰值，我们可认为该峰值很大程度上是由合集的加入造成的，进而可体现合集出现更加频繁的分区。**可知时尚区、娱乐区合集使用更频繁，而舞蹈区、鬼畜区的博主平均使用合集频率更低。**

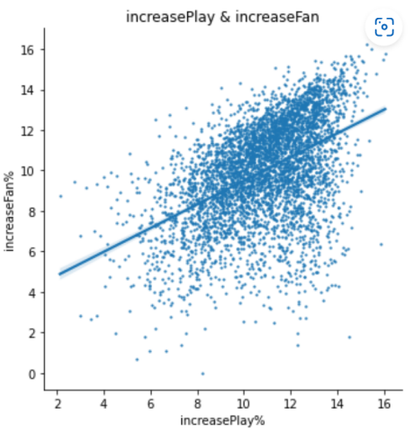


2.2.6 up主视频播放量变化

将过去四年的up主的视频平均播放量做对比，如左图所示。橙色点代表当前各up主的近期平均视频平均播放量，蓝点代表2019平均播放量（取对数）。读图可得当前数据较2019年数据整体有明显提升。因为为随机抽样，我们将其归因于**哔哩哔哩的近三年的发展中用户基数的迅速扩大导致的up主视频的播放量的整体抬升。**

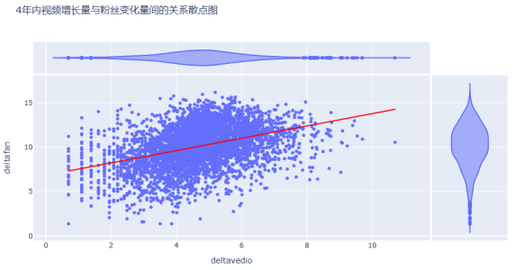
然后我们将过去三年单个up主粉丝数的增加量和视频平均播放增量之间的关系（均取对数）。由右图可得，**粉丝数的增长和用户的视频播放量的增长呈现出正比的关系。**

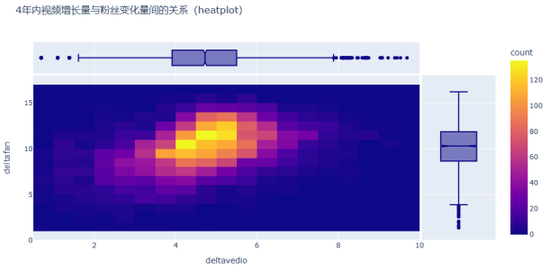




2.2.7 up主视频增量与增粉量的关系

根据新旧数据，我们获取了4年内视频增量与粉丝增量的数据。剔除掉视频增量或增分量为负数的数据（认为该部分数据对up建议的参考价值不大），并对两组数据都取对数后，我们绘制出以下的散点图和热量图：





从散点图趋势线可以看出增粉量随视频增量呈正增长趋势，某种程度上说明**视频产出越多，越容易收获关注。**

从热量图可以看出，数据集中大多数up主（不考虑增粉量、视频增量负增长的up主）在4年内视频增量聚集在在对数坐标下为4-5处（即平均一年更新约14-37个视频），获得约3000-440000个粉丝的增长（可认为是较可观的增长量）。由此可推论，仅从视频更新数量的角度考虑，**up主保持一年14-37个的视频更新量，最有可能获得更多的粉丝关注。**

03 总结与分析

3.1 数据总结

综合以上分析，我们得出了以下结论：

1. up主近期视频的平均播放量、up主的粉丝数的变化量在整体上都呈现出明显的上升趋势，平台的活跃度越来越高，表明哔哩哔哩公司在过去的4年间的发展仍为明显的上升趋势。
2. 从事哔哩哔哩up主这一职业的人群中，男性的比例相对较高，大约为女性数量的两倍左右。不同分区up主男女比例差异明显。
3. 在个人标签上，与“游戏”、”生活“、“音乐”、”动漫“相关的词汇是up主标签中最容易出现的词汇；而在哔哩哔哩的分区上，我们无论是精英数据还是随机数据，我们都能看到将近一半的up主都集中在”游戏“和”生活“两个分区上，”动漫“和”音乐”主题紧随其后，表现了B站虽然正在经历大众化的转型，原初的ACG生态仍有所保留，留住B站“原住民”的同时，还在不断鼓励ACG内容的创作；同时更多生活话题的出现对扩大B站用户基数有着积极作用。
4. 在各个分区上，通过横向对比，我们得出了不同分区up主的“内卷程度”的差异：数据表明，在游戏、卡通、娱乐等分区的内卷程度严重、而舞蹈、时尚、生活的内卷程度较低。将up主以粉丝数是否大于十万为分水岭，进一步筛选出精英up主。在排名的变化上，我们也能够得出相似的结论（2.1.7）。同时，贴上标签、把主题做得相对集中，一定程度上有利于粉丝数的增长。

进一步纵向对比数据，我们得出了各个区的潜在发展潜力。当前“电影”、“生活”、“其它”、“游戏”分区的发展潜力较大，而“娱乐”、“时尚”、“音乐”、“舞蹈”的发展潜力较小。发展潜力体现的是不同分区增粉量（也就是关注人群）的变化情况。虽然数据整体都是扩大增长，但更多的人群开始在“生活”、“其它”，“电影”，“娱乐”领域扎堆，反映出B站文化生态正在实现转型。

5. 在平台关联度上，与游戏、动画、鬼畜区up主与QQ关联度较大；音乐、舞蹈、生活、影视、娱乐与微博的关联度较大；生活、科技、影视、时尚和微信的关联度较大。一定程度上是由于用户的年龄阶段和兴趣方向导致的。QQ的活跃用户主要以年轻人为主，强调用户的个性表达；微信的活跃用户人均年龄较大，更讲究信息传递的有效性。而微博作为大众媒体，能够有效地传递实质信息。平台的属性导致了以上主题关联的数据差异。而在所有区中，只有极少数up主关联了抖音、小红书、淘宝，说明此类app与bilibili平台关联度相对较低。从整体来看，更多的关联号意味着更多的粉丝增长数。推测原因为不同平台的相互关联与吸引增强了up主的影响力。
6. 在up主的活跃度上。我们发现有将近1/3的up主已经不再活跃，其它的up主在一年内仍以不同的活跃度出现在B站上，有近一半的up主依旧保持较高的活跃度（月更或更短）。表明B站的up主创作激励计划等鼓励创作的机制能较有效地增加up主在B站的驻留率
7. 在粉丝数的增量百分比上，我们发现日更的up主与月更的up主的增量中位数基本保持一致，而月更的在`75 percentile`的位置的增量甚至高于日更up主。但从数据整体来看，发布视频数量、发布的总的视频时长是与粉丝量呈正比关系的。进一步表明在提升视频数量的同时，up主也要注意发布视频的质量。才能更有助于吸粉。
8. 从播放视频时长来看，2019年和2022年视频时长差异不明显，表面B站生态在一定程度上对短视频浪潮起到了缓冲作用。

3.2 报告分析

3.1.1 优点

1. 我们利用Web Scraper或取了同样的up主最新的数据，保证了数据的时效性。同时将新老数据对比，得到了关于up主的信息变化的数据并得出了有效结论；
2. 我们的数据的信息覆盖面较广，所有数据均被得到有效利用并得出相关结论；
3. 在一些数据处理上，我们对数据进行分类、加权、整合、以增强结论的有效性，同时数据绘制成不同的类型的数据图，增强了数据的可视化程度。

3.1.2 待改进

1. 在数据获取上，我们获取了合集的数据并将其算入了视频播放量中。对视频播放长度的预测可能会因此带来一系列误差。
2. 在数据权重的配比上，我们采用的是近似一维的线性配比（如0.4，0.3，0.2，0.1）而没有采用更科学的权重配比公式对权重进行赋值
3. 在探究精英阶层和大众up主的主题分区后，我们没有对精英阶层（粉丝数大于十万）的数据做进一步分析，得到精英和普通up主之间的各个其它量之间的差异。