B 站“低创”视频创作者识别

狗熊会 | 精品案例　主讲人：灰灰

目 录

1. 背景介绍

1 发展历程

bilibili

推出创作者激励计划，市值达41.7 亿美元

2018年上市

于2010年由mikufan更名为bilibili

2009年诞生

2014年转型

购买了第一部正版番剧《浦安铁筋家族》

2020年二季度

月活用户(MAU)达到了1.72亿，同比增长55%。市值已达164.16亿美元。

1 激励计划

视频质量和内容升级将是推动 B 站用户增长的主要动力。为此B站于2018年推出了【创作者激励计划】

申请条件

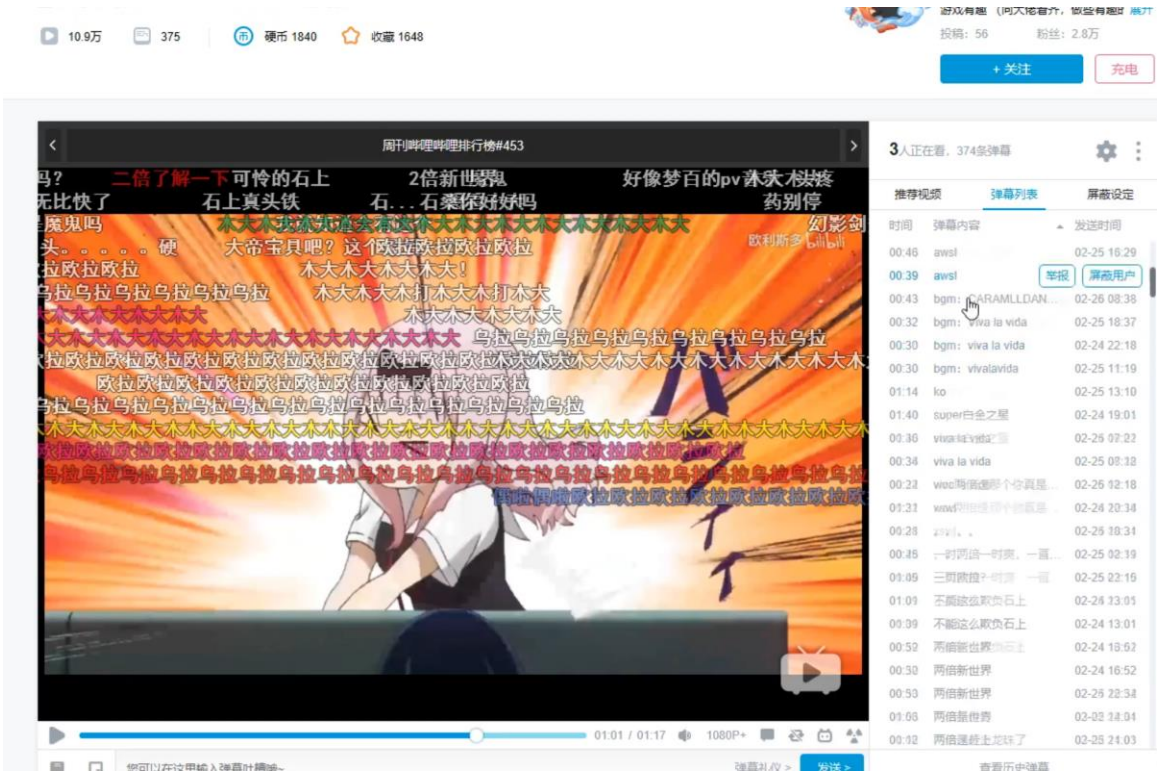
B 站中拥有 1500 个粉丝以上或视频播放总量 15 万以上的用户可以申请加入激励计划。加入之后 B 站会根据之后发布视频的播放量、点赞、收藏、评论等数据给予相应的激励金。

“低创”用户

激励计划引发问题：站内充斥了大量的“低创”用户，减低了B站整体视频质量。

“低创”用户：通过较低的成本，快速生产低质量视频或使用视频标题噱头来获取播放量。

例：20秒原视频后便重复一个画面，但却获得了10万的点击量，连弹幕都有造假嫌疑



低成本
低质量

噱头

1 研究目标

“低创”危害

- 降低B站视频质量
- 高质量的用户生存更为困难
- 激励计划不能得到预期效果

研究目标

- 对B站用户进行群体划分，识别“低创”用户
- 对不同特点的用户提出不同激励计划，做到“精准激励”

2. 数据说明

2 数据说明

变量类型	变量名称	详细说明	取值范围
用户基本信息	性别	定性变量 （3 水平）	男/女/保密
	生日	定性变量（12 水平）	1 月—12 月
	等级	定性变量 （6 水平）	1/2/3/4/5/6
	种类	定性变量 （2 水平）	会员/非会员
用户作品信息	视频数量	定量变量（单位：个）	0 ~ 10000
	音频数量	定量变量（单位：个）	0 ~ 10000
	文章数量	定量变量（单位：个）	0 ~ 10000
	总作品数量	定量变量（单位：个）	0 ~ 20000
用户欢迎度信息	粉丝人数	定量变量（单位：个）	$0 \sim 4 \times 106$
	关注人数	定量变量（单位：个）	0 ~ 2000
	获赞数量	定量变量（单位：个）	$0 \sim 1.2 \times 10^7$
	总播放数量	定量变量（单位：次）	$0 \sim 1 \times 10^9$
	平均播放量	定量变量	0 ~ 600
	获赞率	定量变量	0 ~ 300
	粉丝留存率	定量变量	$0 \sim 1.2 \times 10^7$

数据来源：

B 站的用户信息网站

https://space.bilibili.com

样本数量：

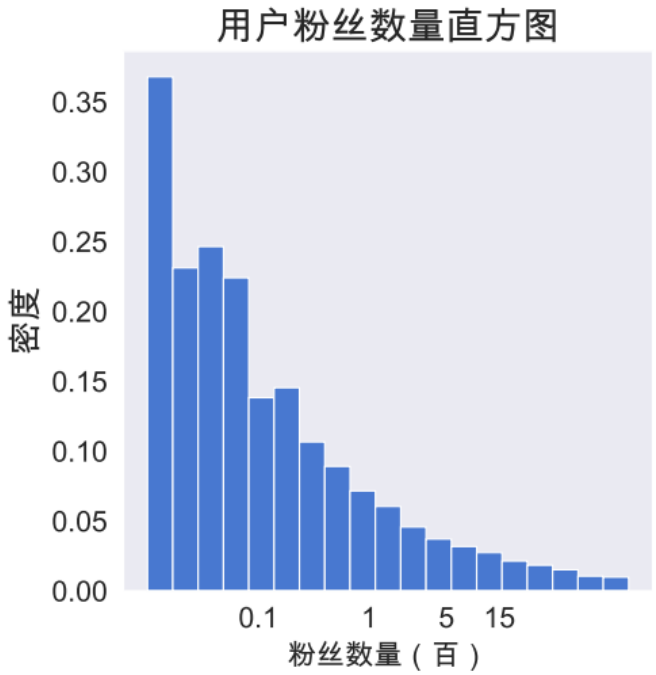
35,157条用户信息

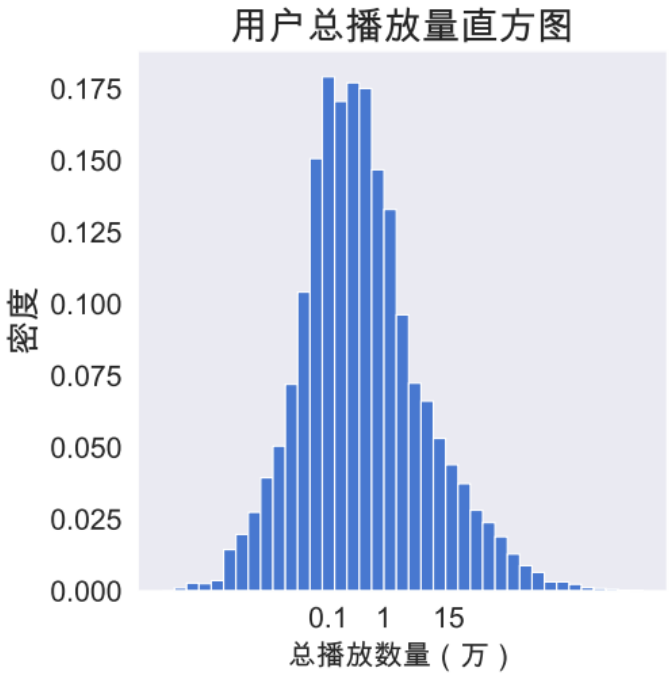
变量概况：

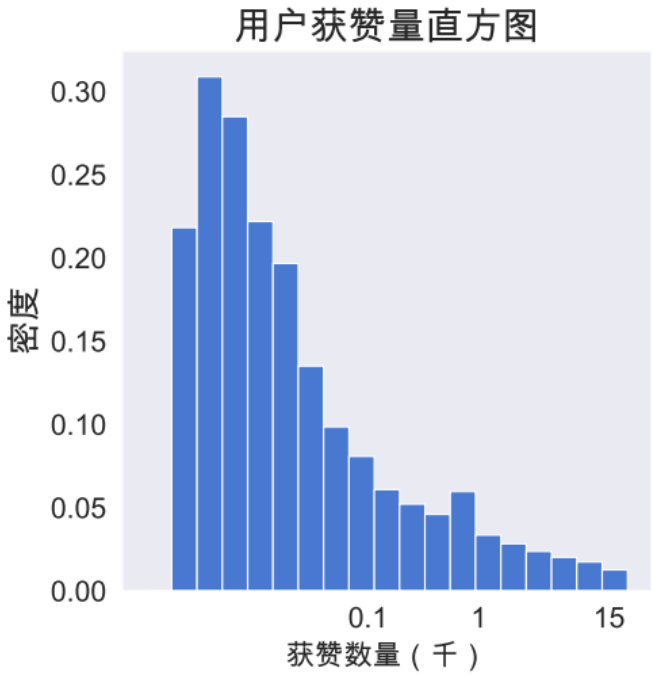
共有16个，包括用户的基本信息，用户作品信息和用户欢迎度信息。

3. 描述分析

3 用户欢迎度信息







- **对数变换后，用户粉丝数和获赞数仍然呈现明显的右偏分布，用户播放量接近正态分布**
- **用户欢迎度之间的差别较大，尤其是粉丝数和获赞数，能够占据榜首的B站博主非常少**

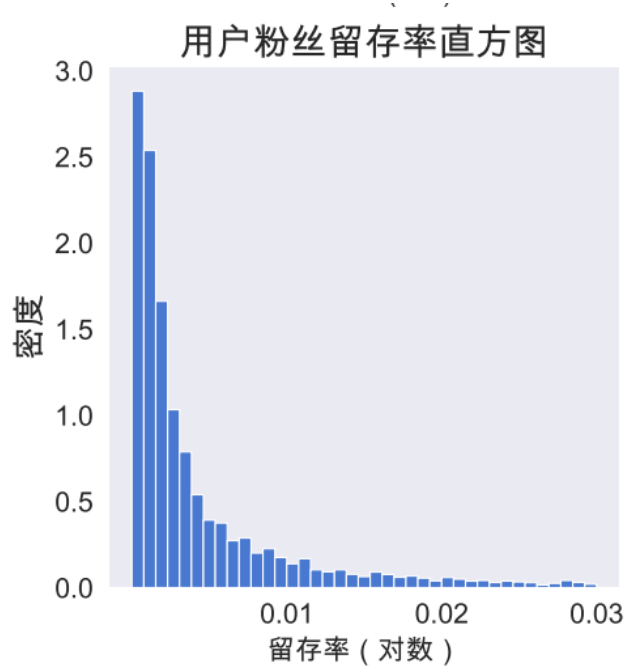
3 用户欢迎度信息

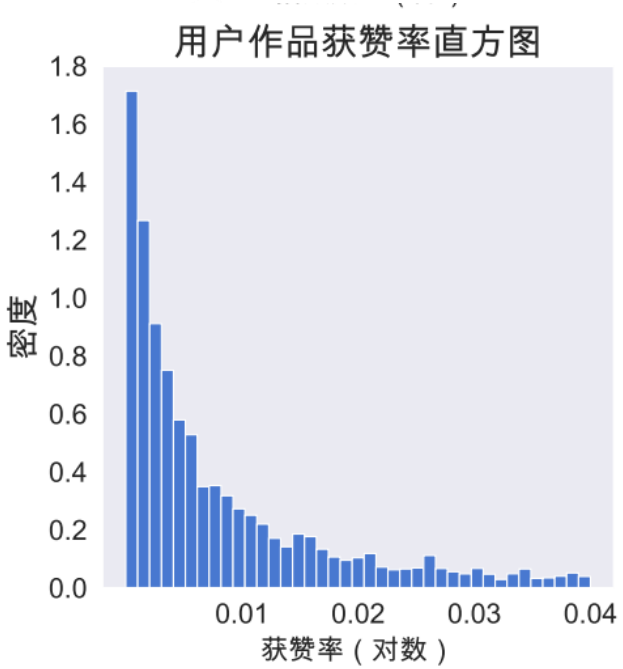
粉丝留存率 = 粉丝数/视频播放量

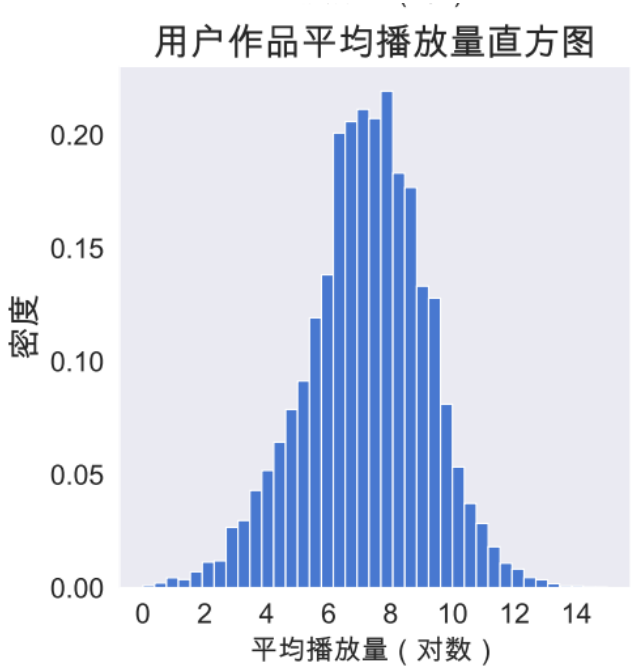
获赞率 = 获赞数/视频播放量

平均播放量 = 作品总播放量/总作品数

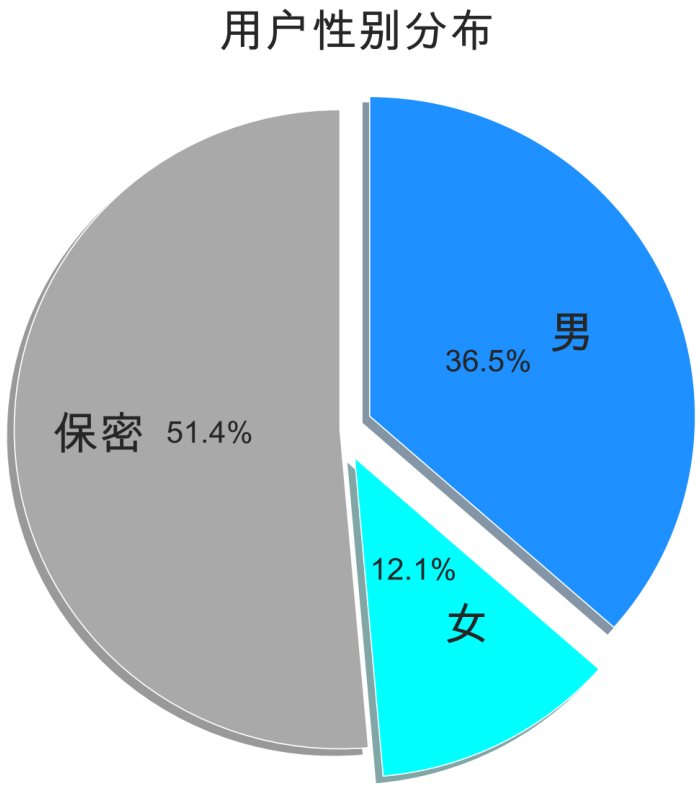
- 对数变换后，用户粉丝留存率和获赞率仍然呈现明显的右偏分布，用户平均播放量接近正态分布
- 综合用户欢迎度信息的六个变量不同分布的特点，说明存在不少用户具有高播放，低留存率，低获赞率的特点。

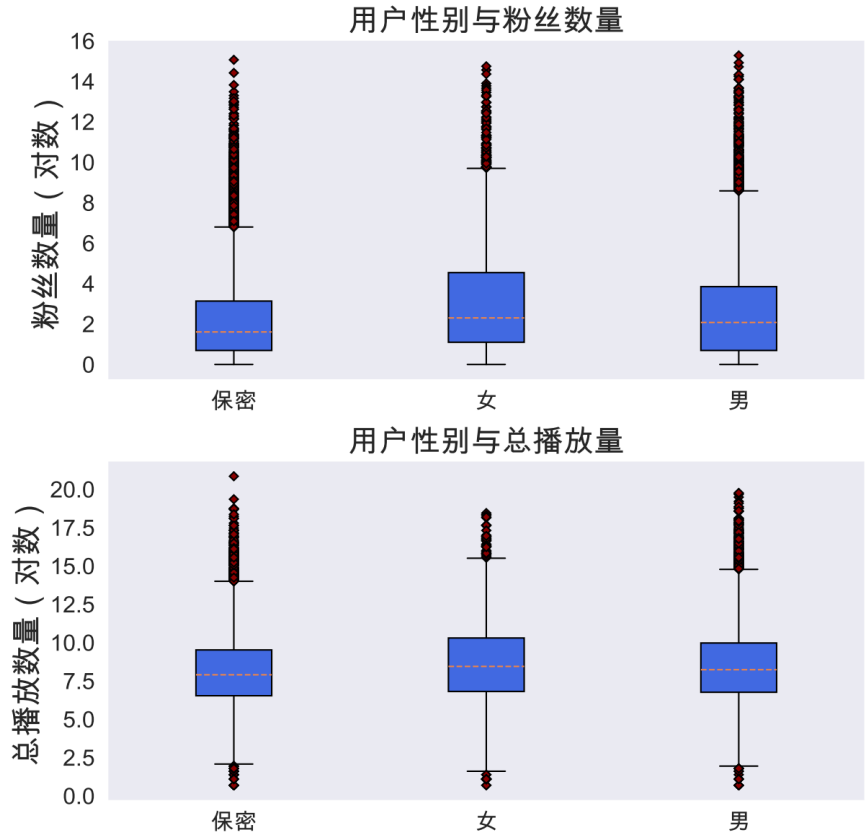






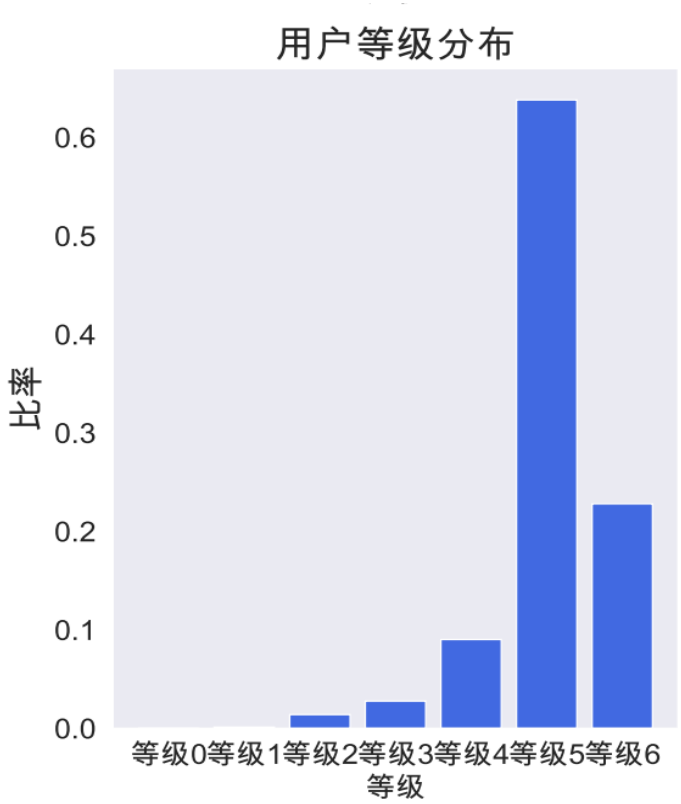
3 用户基本信息

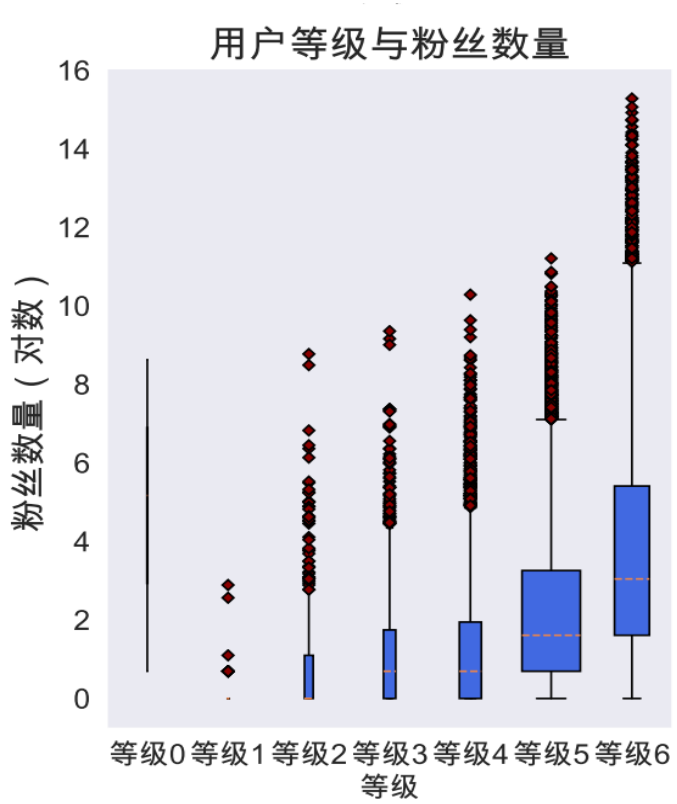


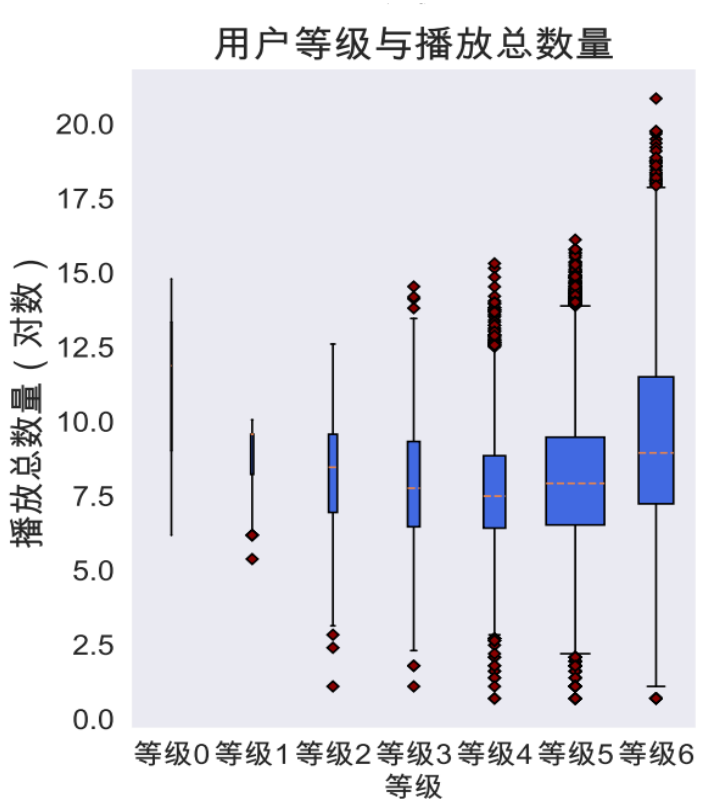


- **男性用户是女性用户的三倍，说明 B 站更受男性欢迎**
- **填写性别信息的用户粉丝数量与播放数量都较高**

3 用户基本信息

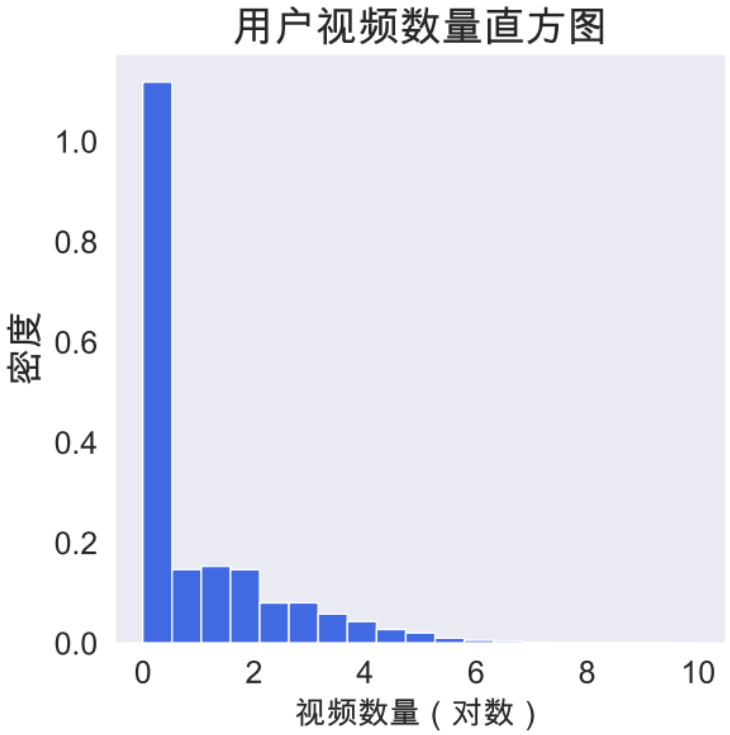


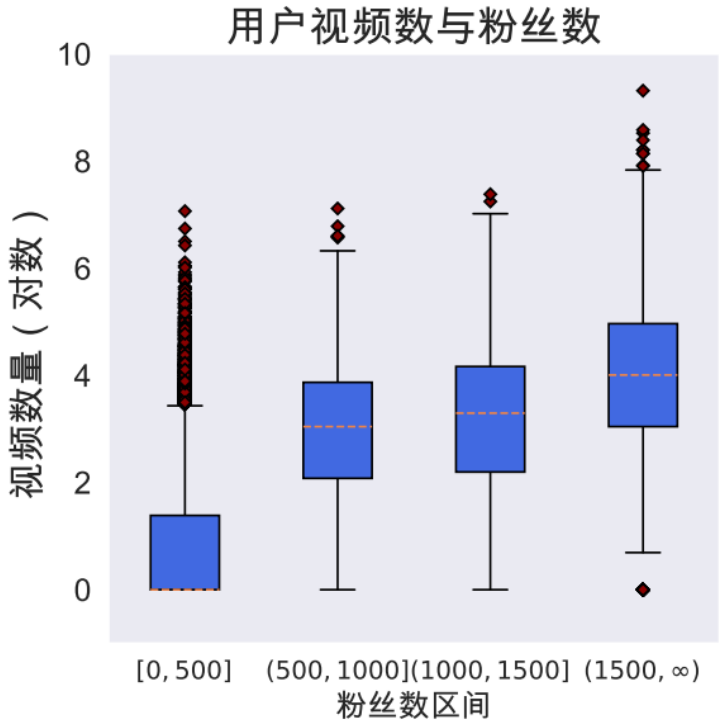


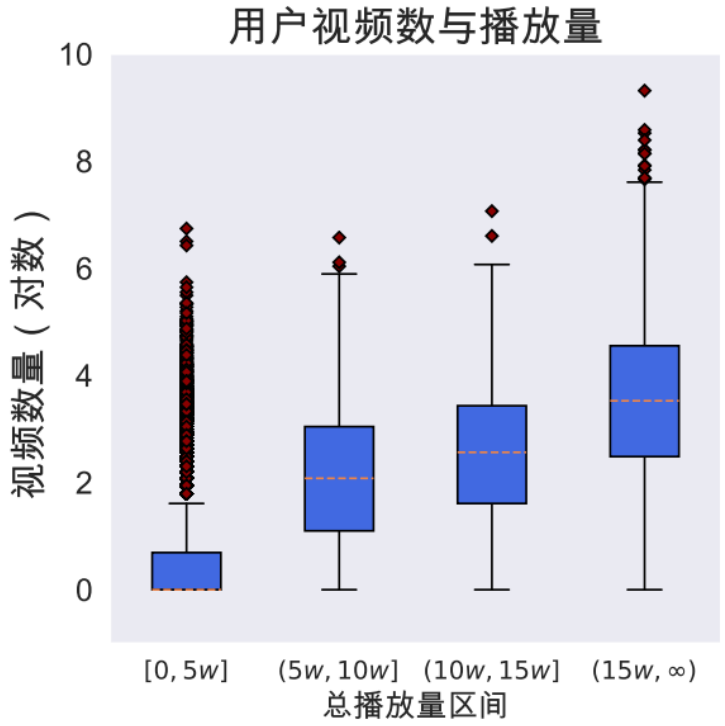


- B 站中的用户等级是根据用户在 B 站的活跃程度进行评级的结果
- 目前样本中等级为 5、 6 的用户数量明显高于其他等级；并且等级越高、粉丝和播放量越多

3 用户作品信息

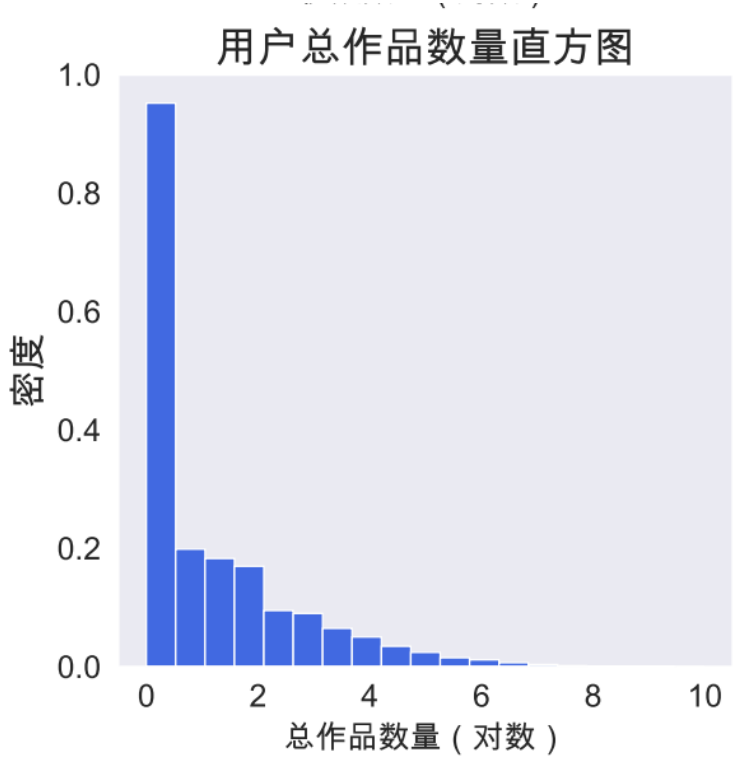


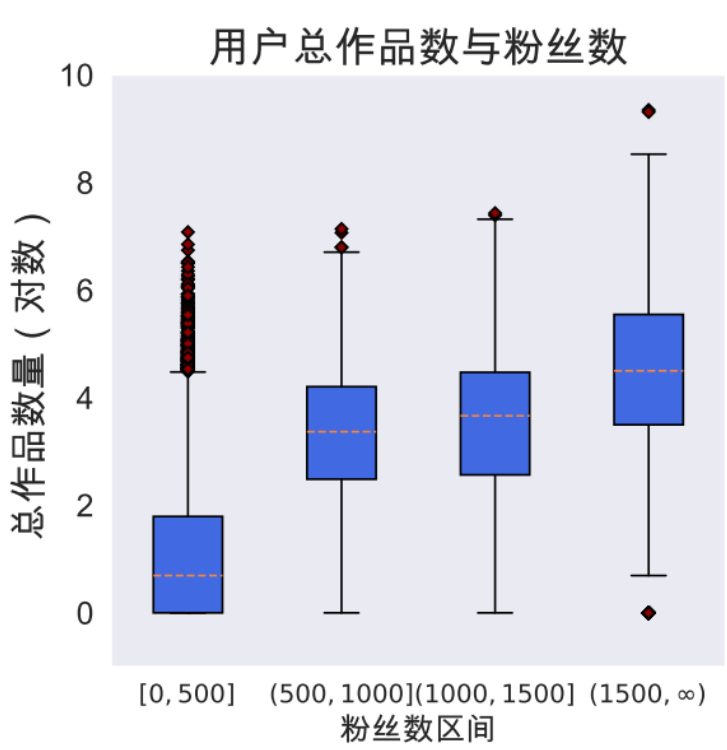


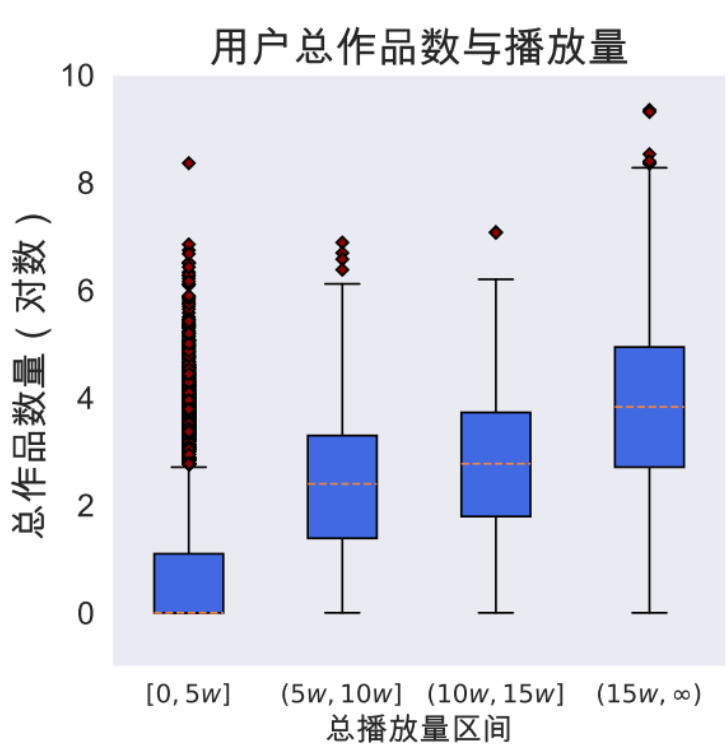


- **用户视频数量呈现明显的右偏分布，拥有大量视频的用户较少**
- **用户视频数越多，其拥有的粉丝数和播放量也相对较高**

3 用户作品信息





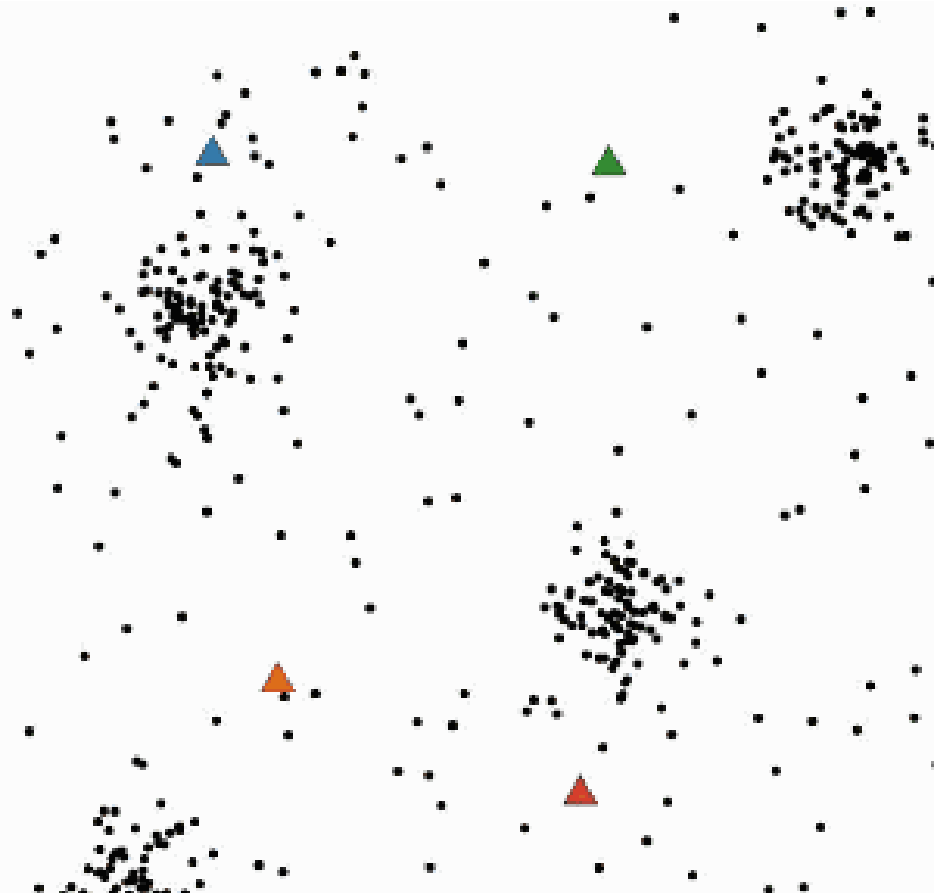


- **用户作品数量呈现明显的右偏分布，拥有大量作品的用户较少**
- **用户作品数越多，其拥有的粉丝数和播放量也相对较高**

4. 聚类分析

4 聚类方法介绍

聚类是一种无监督机器学习方法。给定一组数据点，我们可以使用聚类算法将每个数据点划分到一个特定的组。同一组中的数据点应该尽可能相似，而不同组中的数据点应该尽可能不同。



4 K-prototype聚类

- **K-prototype**是处理混合特征聚类的典型算法，它的核心是对**连续型变量**和**离散型变量**定义不同的**“距离”**指标
- 假设数据集有**n**个样本，每个样本包含**m**个变量，其中前**p**个为数值型变量，后**m-p**个为离散型变量

离散型变量	连续型变量
• **海明威距离** $\delta(x_{ij}, x_{kj}) = \begin{cases} 1, & x_{ij} != x_{kj} \\ 0, & x_{ij} == x_{kj} \end{cases}$	• **欧式距离** $\delta(x_{ij}, x_{kj}) = (x_{ij} - x_{kj})^2$

样本i与聚类中心l的距离： $d(x_i, Q_l) = \sum_{j=1}^{p}(x_{ij} - q_{lj})^2 + u\sum_{j=p+1}^{m}\delta(x_{ij}, q_{lj})$

4 聚类个数的确定

Gap算法

从均匀分布中随机生成n个样本点，对其聚类后，计算组内平方距离，然后与原始样本聚类后的组内平方距离做差，差值越大说明聚类效果越好

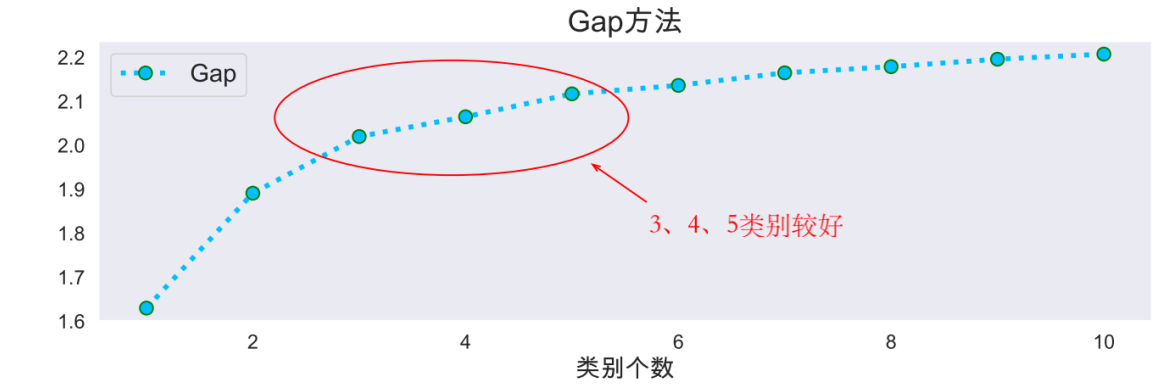
肘方法

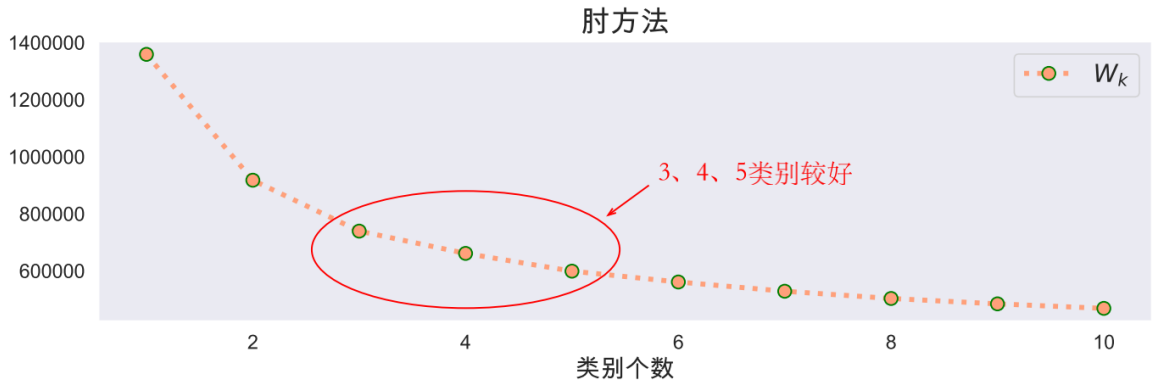
在一定的类别个数下，计算组内各样本之间的平方距离，距离变化剧烈的类别为最优类别。

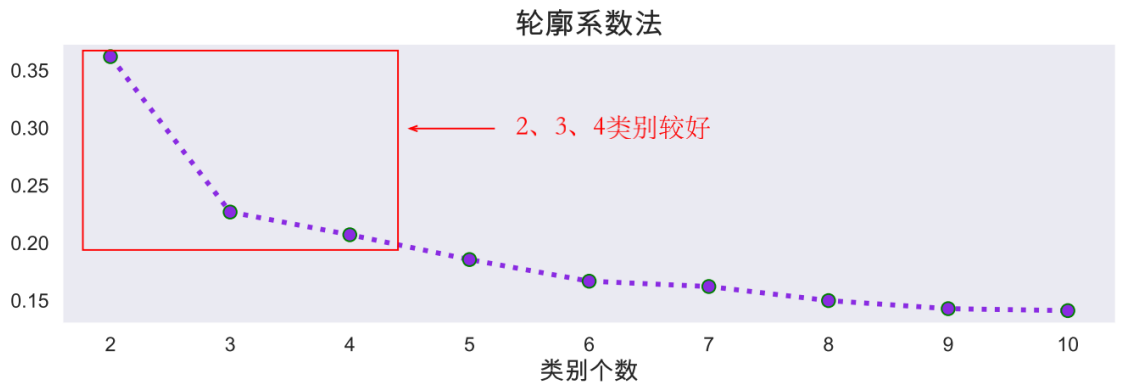
轮廓系数法

对比：**1.** 样本与其他组样本的距离， **2.** 样本与自己组内样本的距离，做差后代表轮廓系数，越小表示聚类效果越好

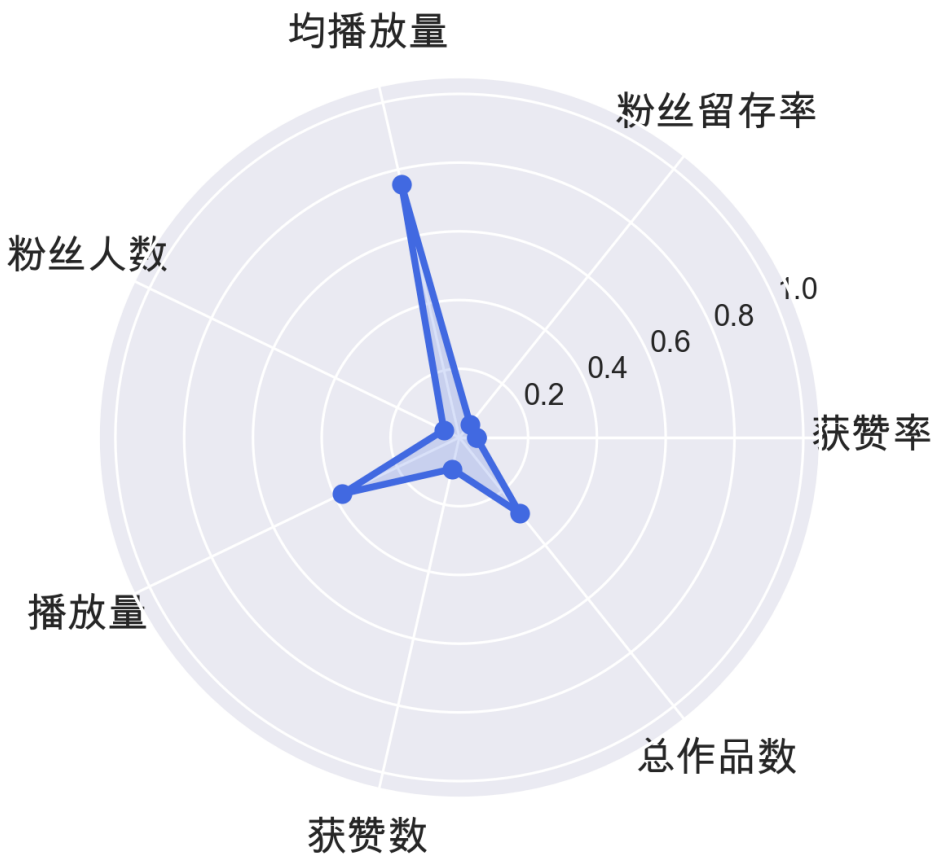
综合来看选择4类进行聚类







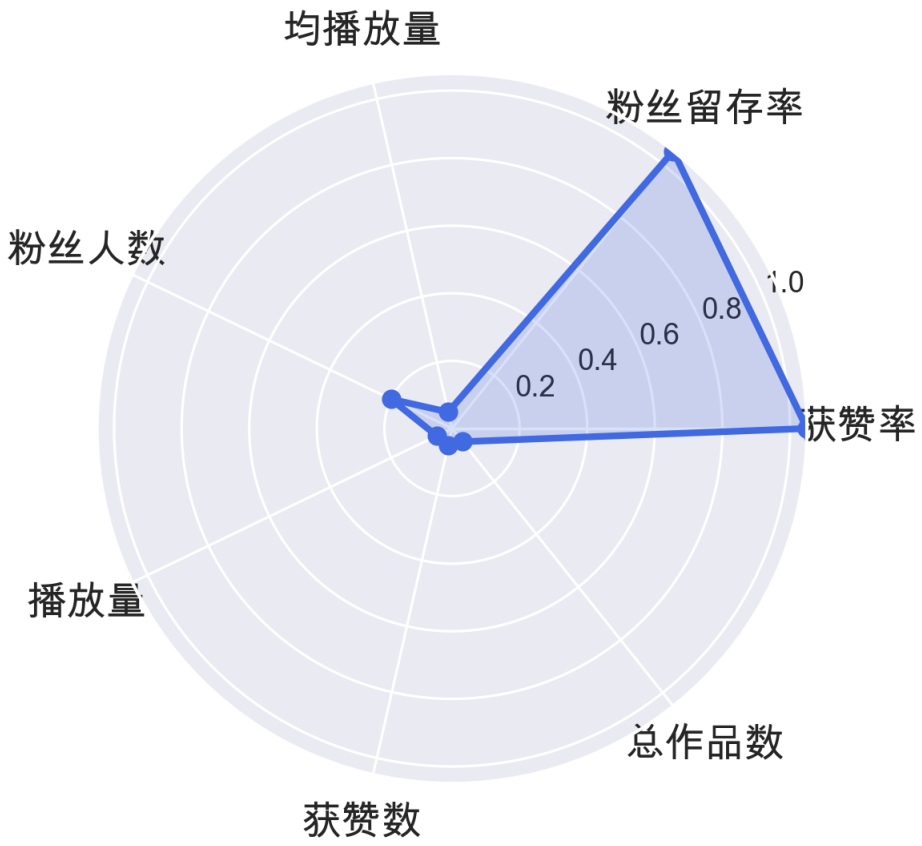
4 聚类结果



1 “低创用”户，占比15.2%

第一类用户呈现高播放量，低粉丝数，低获赞率，低粉丝留存率的特点，符合之前分析的大部分“低创”用户的特点，其粉丝留存率与获赞率并不高，将这一类用户称为**“低创“用户。**

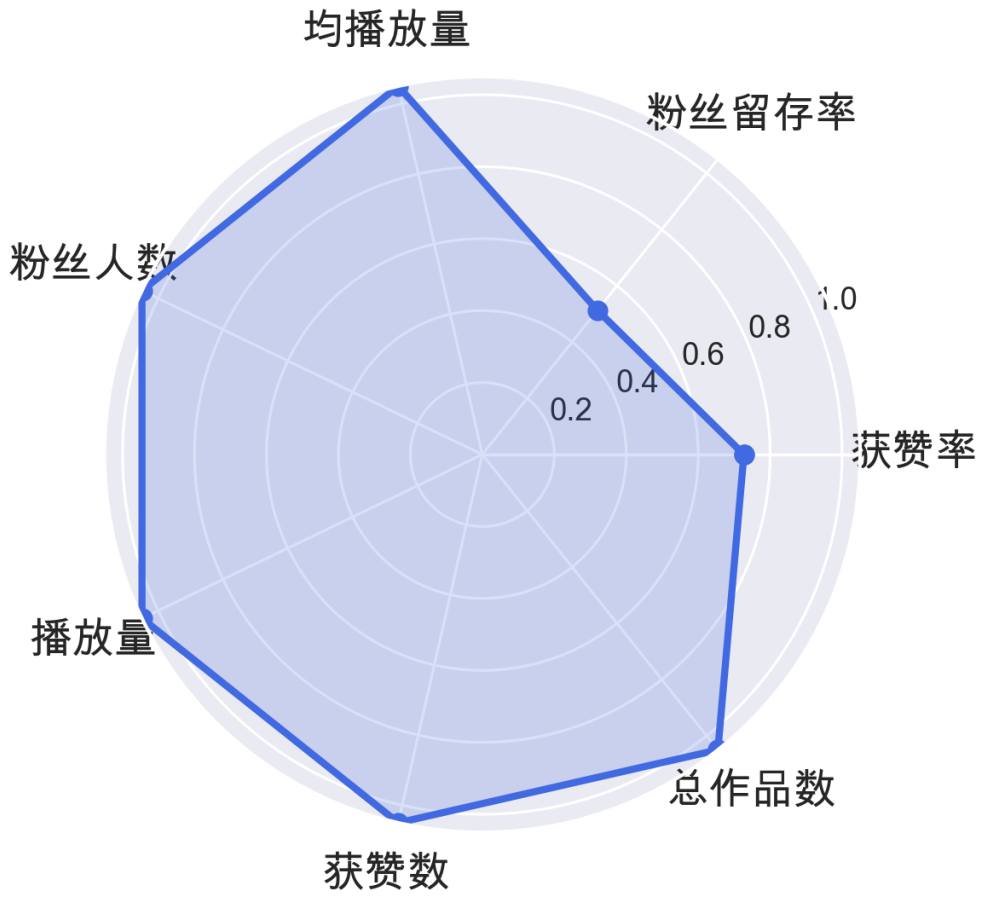
4 聚类结果



2 高质量用户，占比29.3%

第二类用户是具有低播放量，高粉丝量，高获赞率，高粉丝留存率的特点，这说明该类用户创作的作品质量高，能够用较少的播放量即可吸引较多的粉丝，将这一类用户称为**高质量用户。**

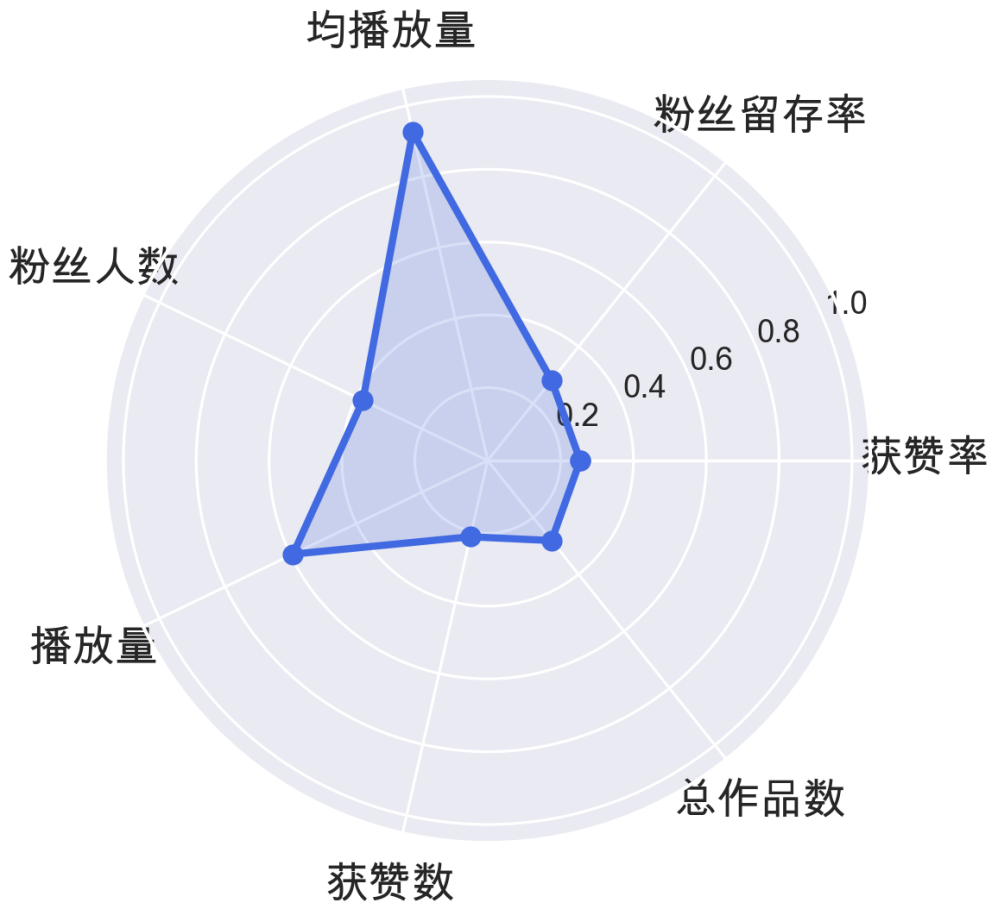
4 聚类结果



3 up主用户，占比 23.7%

第三类用户具有超高播放量，超高粉丝量。其作品数量较多，质量也呈现了参差不齐的情况，因而这一类用户在获赞率与粉丝留存率上并不一定很高，将这一类用户称为 **up 主用户。**

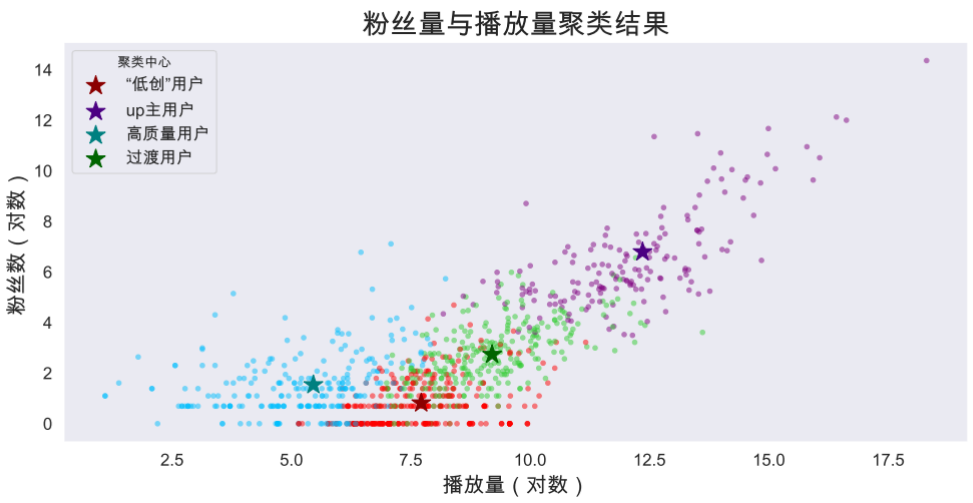
4 聚类结果

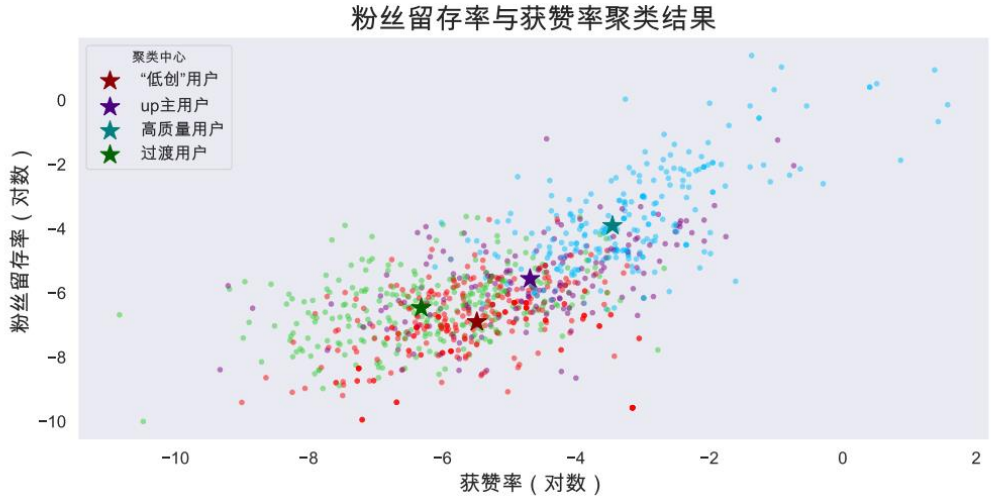


4 过渡用户，占比31.8%

第四类用户介于三者之间，构成较为复杂，实际上是一种过渡的用户，不适于对这类用户进行简单地归类，这里将这一类用户称为**过渡用户。**

4 各类用户交叉分析





- 低创用户视频播放量较于过渡用户以及 **up** 主用户低，比高质量用户高很多
- 高质量用户较其他用户，明显是低播放量，高获赞率，高粉丝留存率特点
- 过渡用户在粉丝量与播放量上是最接近 **up** 主用户的，但在获赞率与粉丝留存率上其与低创用户较为相似这说明羁绊过渡用户成为 **up** 主的**原因很有可能是视频质量低，留不住粉丝。**

5.应用与结论

5 商业应用

监测识别

通过聚类模型，识别“低创”用户，对于粉丝留存率、获赞率过低但仍产生大量视频的用户，适当推荐其参加B站所推出的一些激励活动，或删除其部分“低创”视频，以提高B站整体视频质量。

激励措施

提高“低创”用户在激励计划中的门槛，从而激励这部分用户，通过提高其作品质量而进入激励计划，而不是仅仅只提高其作品播放量以及作品数量。

5 商业应用

根据其他不同用户特点，提出对应的的激励计划：

- 对于**高质量用户**低播放量、较低粉丝量、高粉丝留存、高获赞的特点，适当将激励计划的粉丝量或播放量阈值降低，特别是视频播量阈值。提高其作品收益，从而创作出更多有价值的作品。

- 对于**过渡类用户**高粉丝以及获赞量、高播放量、低粉丝留存率、低获赞率的特点。应重点关注其视频质量。加强对这部分人视频质量监控，有助于提升 B 站视频质量。

- 对于**up 主用户**，由于up主用户粉丝量与播放量已经相当巨大，不少已经达到了激励计划的标准，可以维持之前的粉丝量与播放量阈值。

5 结论

研究目标

- 识别“低创”用户，解决由激励计划所引发的“低创”问题。

研究方法

- 通过聚类方法对用户群体进行了划分，较好的识别了“低创”用户。

应用

- 根据不同类别用户特点，设定不同类别的激励计划，做到“精准激励”。

启发

- 网站对创作者设置激励类计划，可先对创作者进行分类，根据每类特点设定不同的激励类计划。
- 聚类可作为UGC视频网站识别“低创”用户的一种有效方法。

狗熊会 | 精品案例

扫描二维码，关注狗熊会，获取更多案例资源

谢谢观赏！