



# 哔哩哔哩文化社区模式 成功之道探索

汇报：王储  
时间：2019.09



# 目录

## CONTENTS

- 01 项目简介
- 02 数据统计可视化
- 03 高分视频影响因素分析
- 04 视频评论分析
- 05 总结

01

# 项目简介

# 项目简介

哔哩哔哩现为国内领先的年轻人文化社区，被粉丝们亲切的称为“B站”，它一直以“下一代的文化乐园”和“Z时代社区”这样的形容自居，近年来越发引得众人好奇。同样是做弹幕视频，同样是以年轻人为主，为什么A站在2018年初因资金链断裂黯然闭站，而B站却能够激流勇进在美上市？



# 数据说明

## 数据预处理

通过 Pandas 对爬取的 1200 条视频信息和超过三万条评论信息进行缺失值填充和排序，对数据进行分列处理和数据格式转化



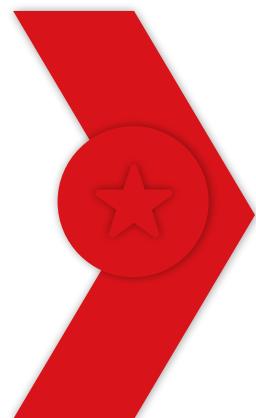
## 相关字段

类别排名, 点赞数, 投币数, 收藏数, 分享数, 评论数, 播放量, 弹幕数, 详细类别, 更新时间, 全站 rank 等



## 建模分析

建立播放量与其他 7 个度量的多元线性回归模型并通过 sklearn 建模预测, 高分视频影响因素 K-Means 聚类模型, 评论字段的情感分析与 LDA 主题建模



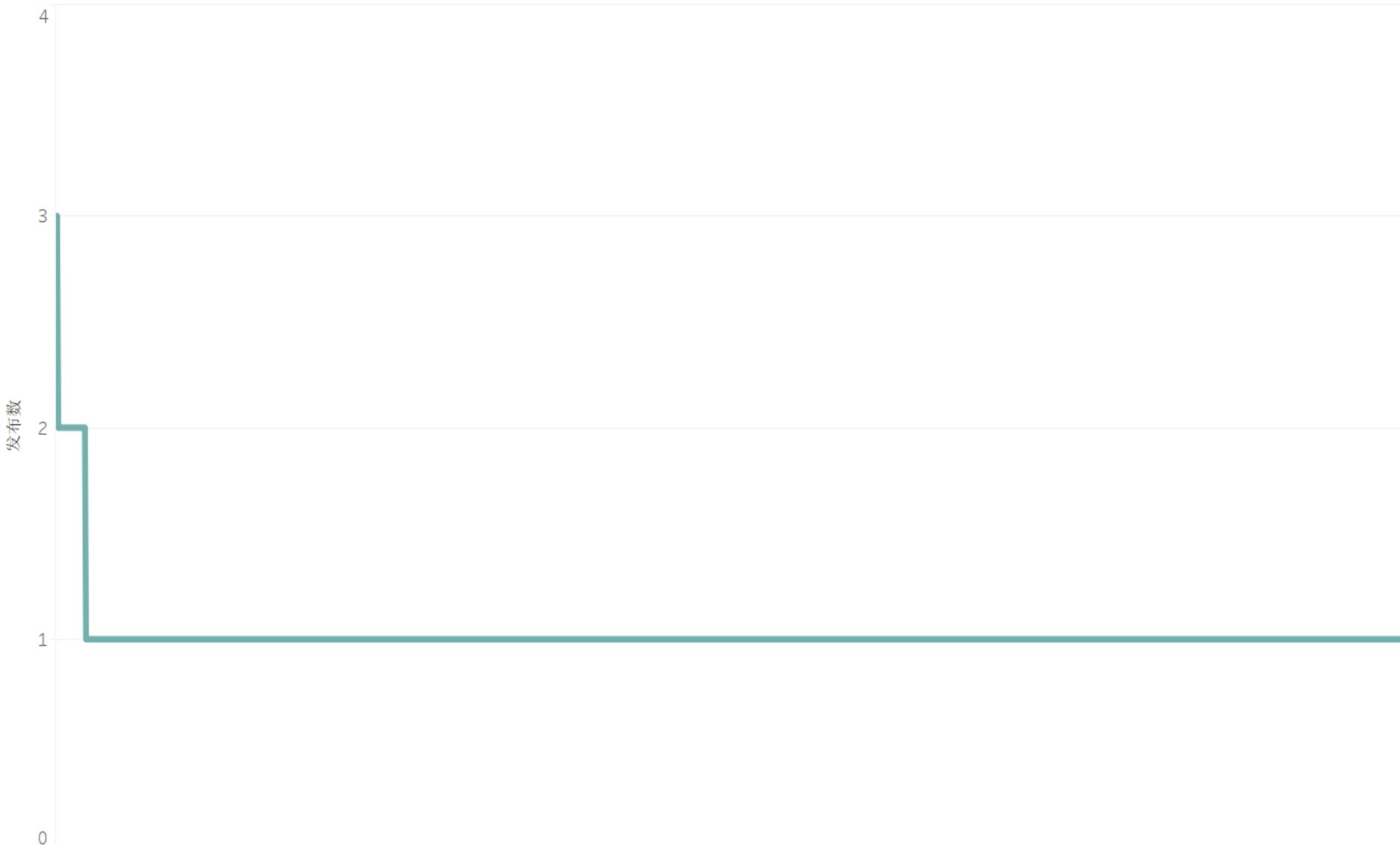
本项目收集了 B 站原创榜的全部数据, 对综合得分、类型、播放量、弹幕、点赞、投币、收藏的等多维度数据进行分析, 将不同类型的视频内容进行对比分析, 研究二次元、游戏、鬼畜谁才是 B 站的头牌。建立高分视频影响因素模型, 洞察 B 站视频内容特色, 研究当下年轻人兴趣所在。

02

# 数据统计可视化

# 总览—作者发布数量分析

作者发布数量分析

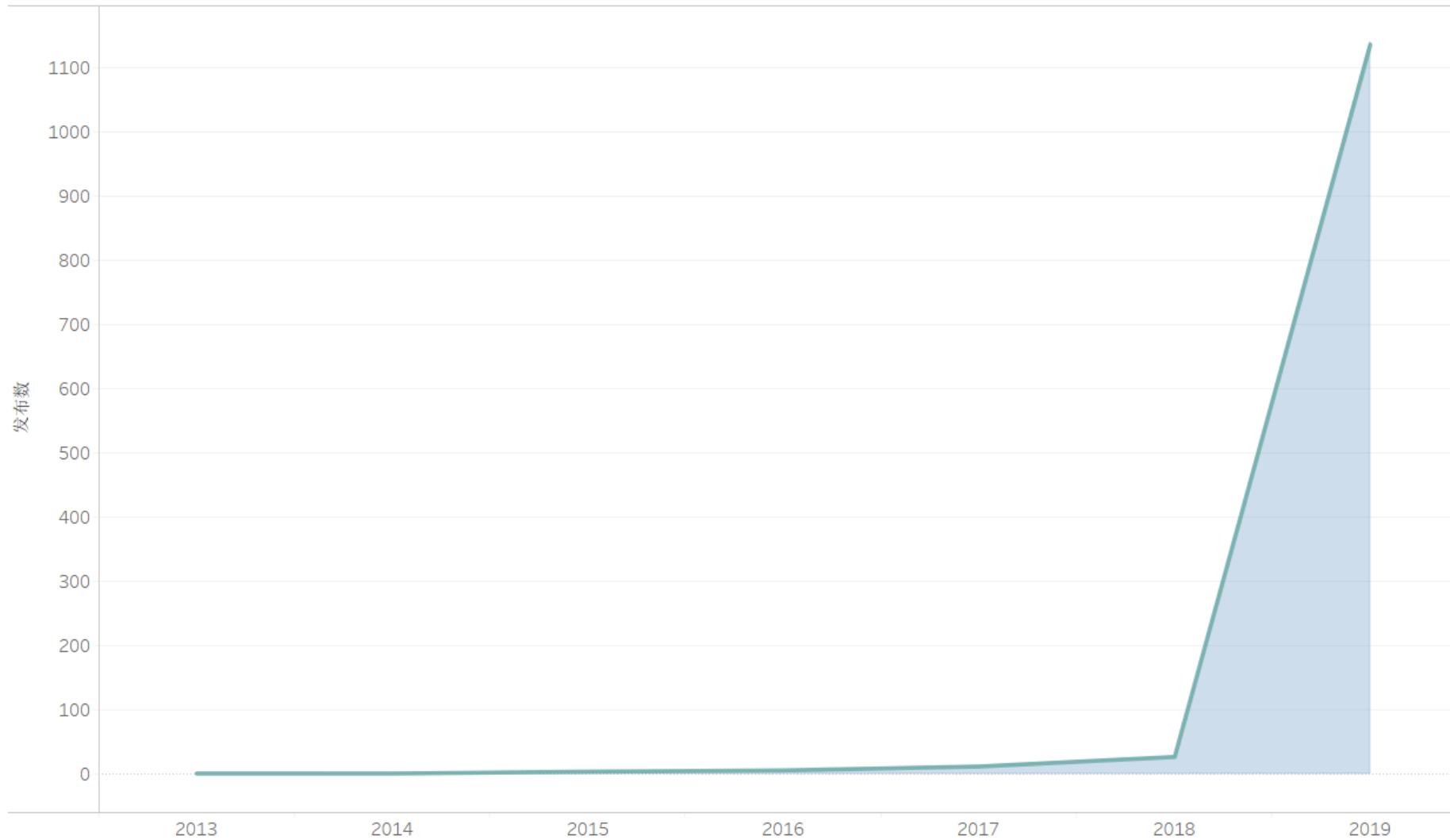


◆ 爬取的高分样本视频共计1200个，剔除缺失值后共计有1170位作者参与发布，有超过1100位作者只产出了1部高分视频，说明了高分视频的偶然性，与特定作者无紧密关联



# 总览一历年视频发布数量

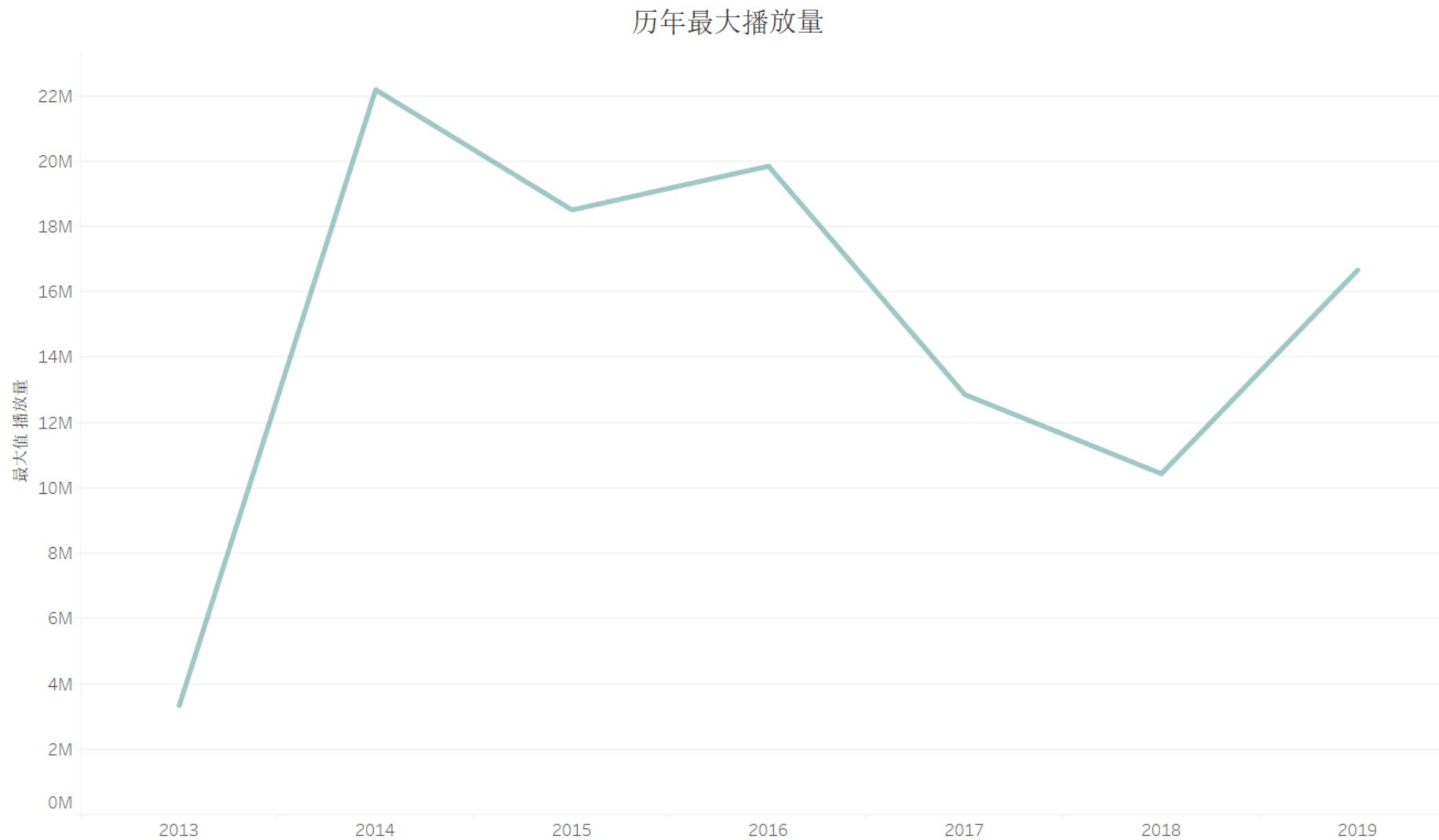
历年发布数量



◆ 超过1100部高分视频均是发布在2019年以后，相比于前五年有很大的飞跃，说明自B站2018年赴美上市后，高分视频数量激增或与视频质量以及发布总体数量上涨有关



# 总览一历年视频最大播放量

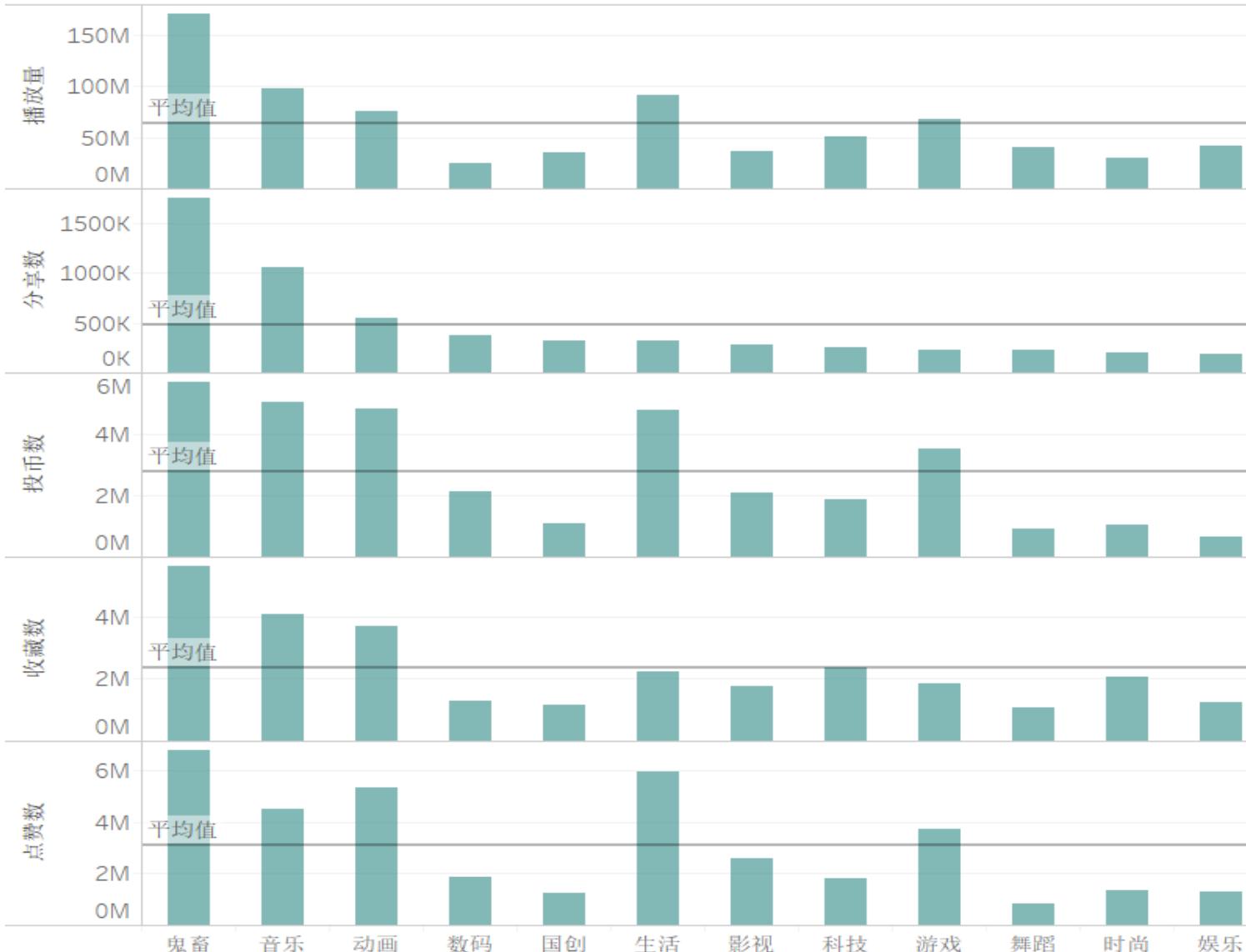


◆ 与发布数量不同，  
最大播放量视频  
出现于2014-2016  
年区间，均显著  
高于2017-2019年，  
侧面反映出2014-  
2016年期间是B站  
视频质量较高但  
产出数量较少的  
发展期间，也为  
后期2018年上市  
打下了可靠的基础



# 不同类别视频基于播放量、投币数、收藏数等维度的可视化分析

## 播放、分享、投币、收藏、点赞数分析

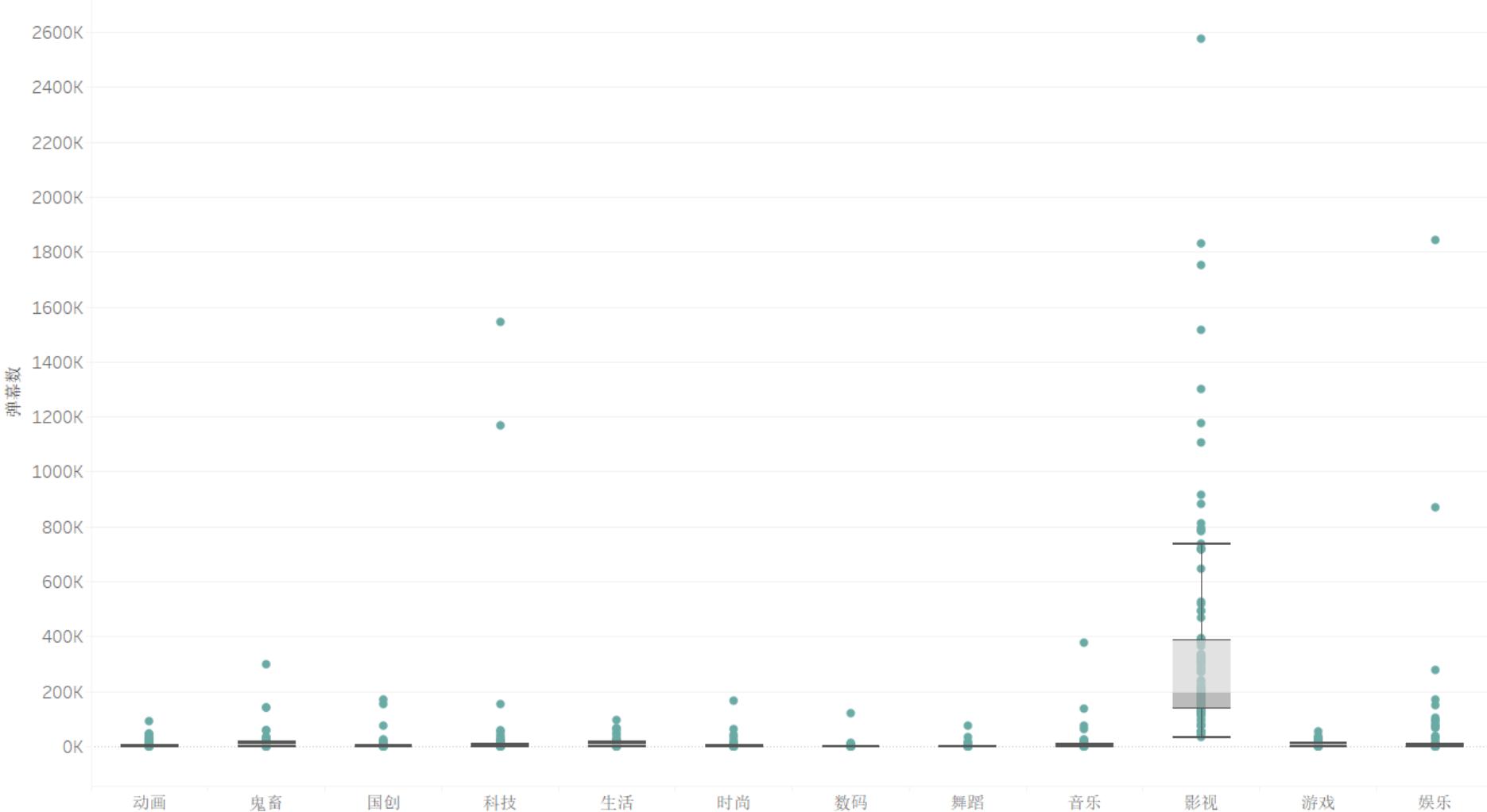


- ◆ 总体来看，鬼畜的五项指标均远超平均值，领先于其他11类视频，成为B站的头牌
- ◆ 音乐、动画、生活类视频在五项指标上几乎均高于平均值
- ◆ 舞蹈、娱乐、国创类视频受欢迎程度显著低于前几类视频



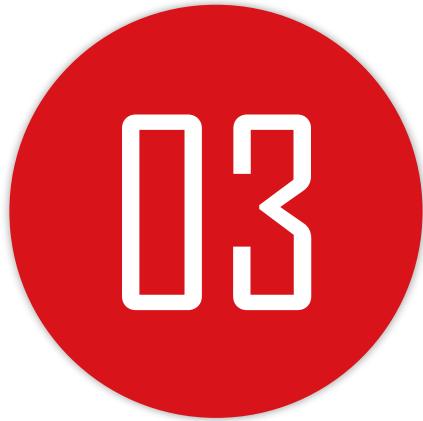
# B站不同类别视频弹幕数分析

不同类别弹幕数盒须图



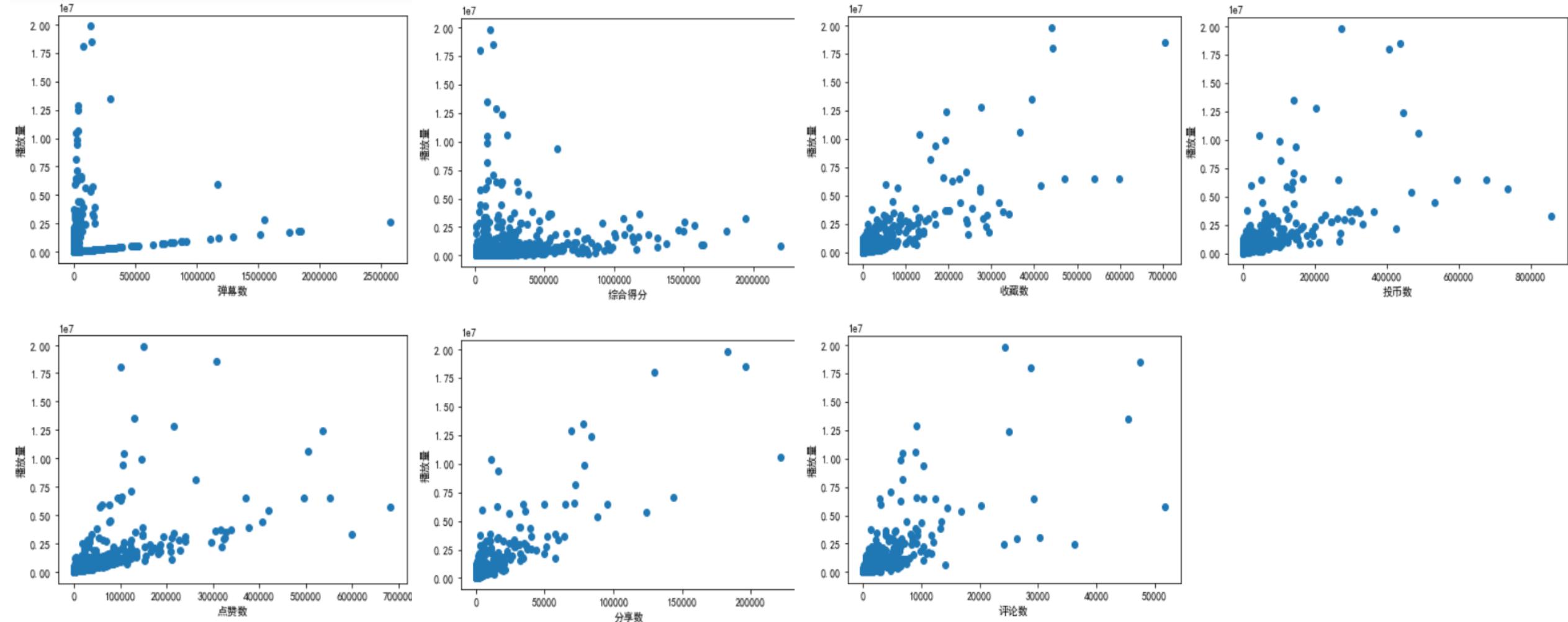
- ◆ 作为B站的最大特色的实时评论功能—弹幕，影视类视频的弹幕总数近千万条，也是弹幕数分布最广的视频类别
- ◆ 除娱乐和科技类视频外，其余类别弹幕数分布无很大差别





# 高分视频影响因素分析

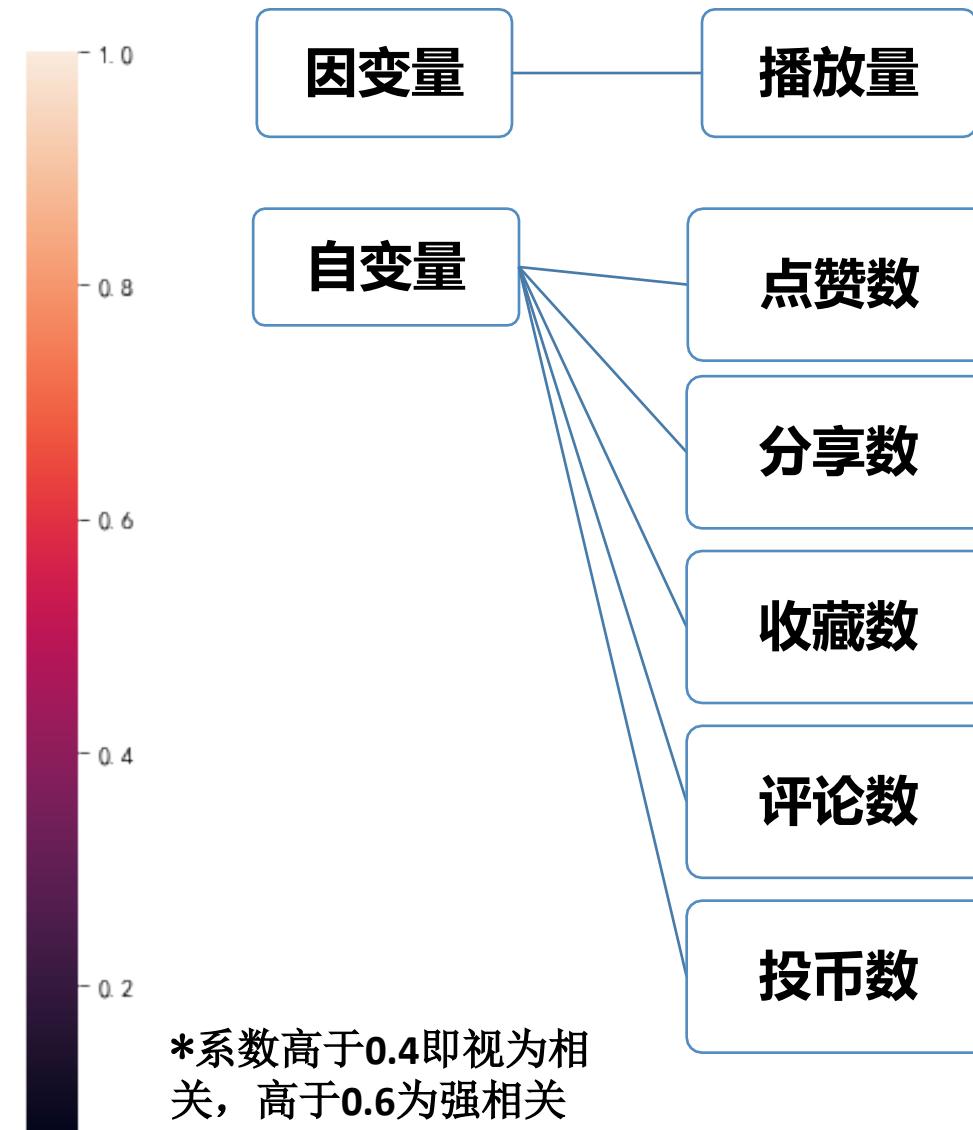
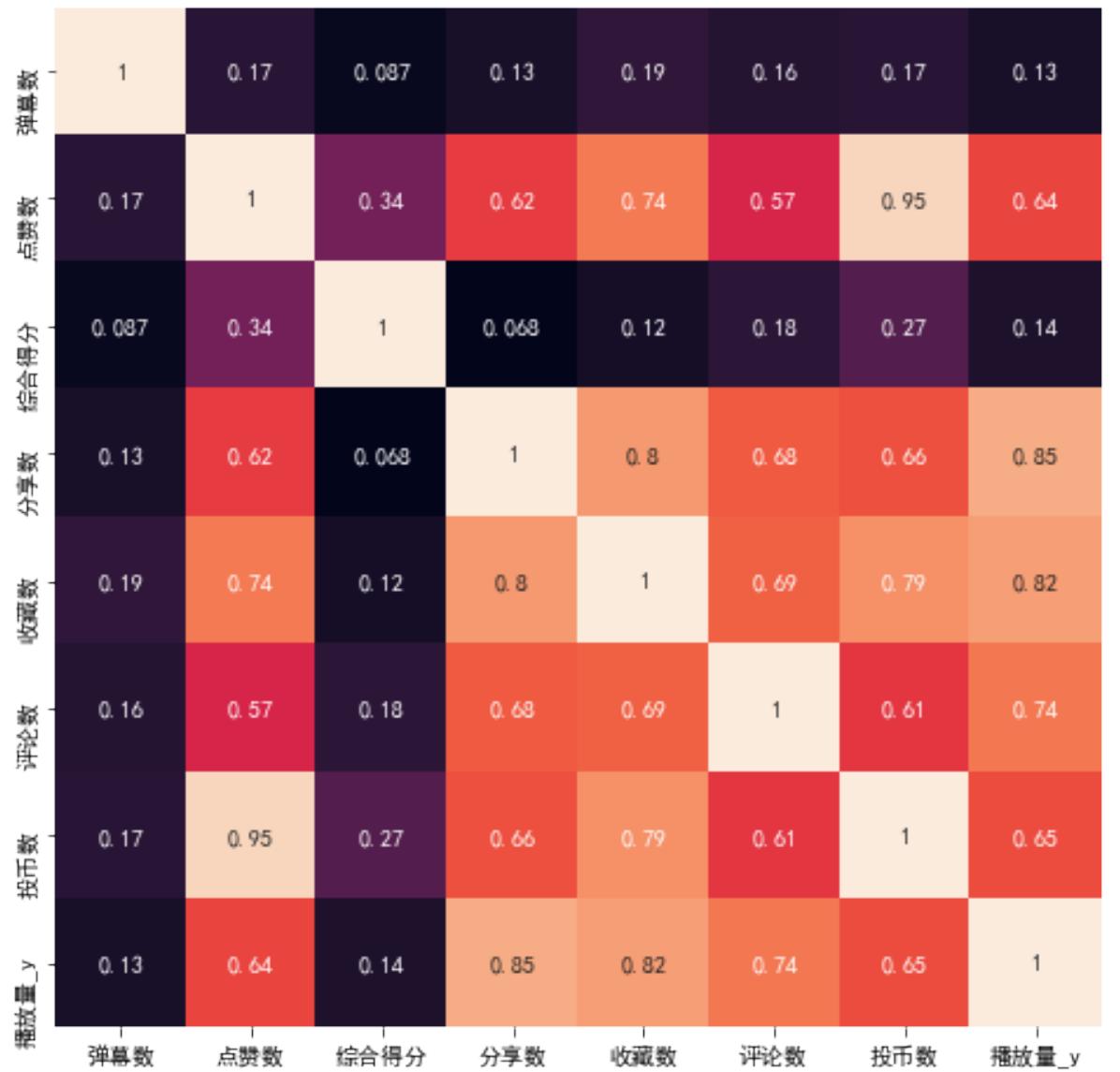
# 多元线性回归分析—播放量与其他维度散点图分析



- ◆ 建立剔除异常值后的播放量与其他维度关系散点图，可以看出除**弹幕数**与**综合得分**与播放量成负相关外，其余维度数据均与播放量成正相关，且可以看出大部分视频播放量均集中在12万次上下



# 数值维度相关热力图分析—探索不同维度之间的相关性影响



# PCA降维后的多元线性回归分析（3个主成分）

```
OLS Regression Results
=====
Dep. Variable: y R-squared: 0.728
Model: OLS Adj. R-squared: 0.727
Method: Least Squares F-statistic: 1594.
Date: Fri, 27 Sep 2019 Prob (F-statistic): 0.00
Time: 14:58:55 Log-Likelihood: -17932.
No. Observations: 1195 AIC: 3.587e+04
Df Residuals: 1192 BIC: 3.589e+04
Df Model: 2
Covariance Type: nonrobust
=====
            coef    std err        t      P>|t|      [0.025      0.975]
const    5.944e+05  2.3e+04   25.787  0.000  5.49e+05  6.4e+05
x1       6.35e+05  1.16e+04   54.840  0.000  6.12e+05  6.58e+05
x2      -3.043e+05  2.26e+04  -13.440  0.000 -3.49e+05 -2.6e+05
=====
Omnibus: 1168.555 Durbin-Watson: 1.885
Prob(Omnibus): 0.000 Jarque-Bera (JB): 127863.532
Skew: 4.229 Prob(JB): 0.00
Kurtosis: 52.964 Cond. No. 1.99
=====
```

基于上述变量建立多元线性回归模型



初步建立的线性回归模型：

💡 存在多重共线性

💡 评分变量不显著



经多次修正：

除综合得分和弹幕数外的所有变量

进行主成分分析后建立修正多元线性回归模型

尽可能提取原始变量信息，消除多重共线性影响

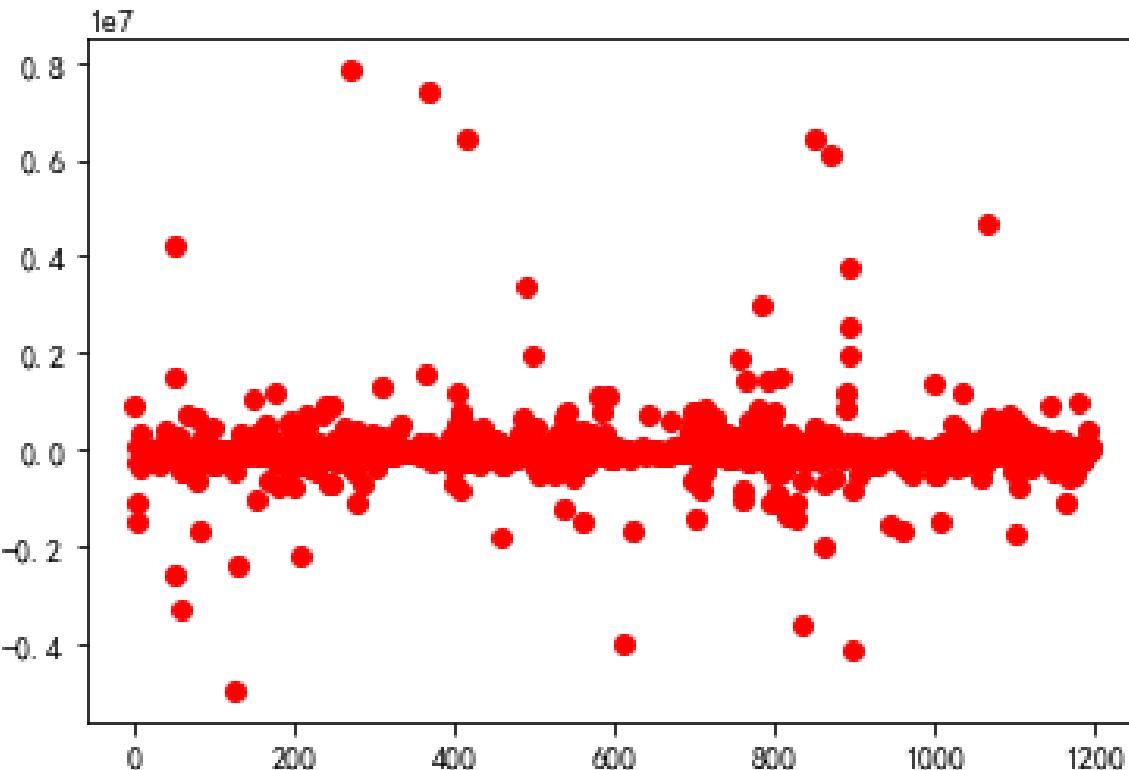
主成分载荷 →

	弹幕数	点赞数	综合得分	分享数	收藏数	评论数	投币数
X1	0.12	0.45	0.15	0.42	0.46	0.4	0.46
X2	0.36	0.15	0.84	-0.29	-0.19	-0.14	0.06

预测播放量  $Y = 5.9 \times 10^5 + X1 \times 6.4 \times 10^5 + X2 \times (-3.04) \times 10^5$



# 岭回归预测模型



- ◆ 通过机器学习模型对1200个视频样本进行学习，绘制出训练集预测模型散点图，从散点图可看出预测值与真实值差值大致稳定在0轴两侧

回归系数: -34247. 331498. 39450. 710124. 499358. 348659. -395252.  
(依次为上述表格中变量顺序)

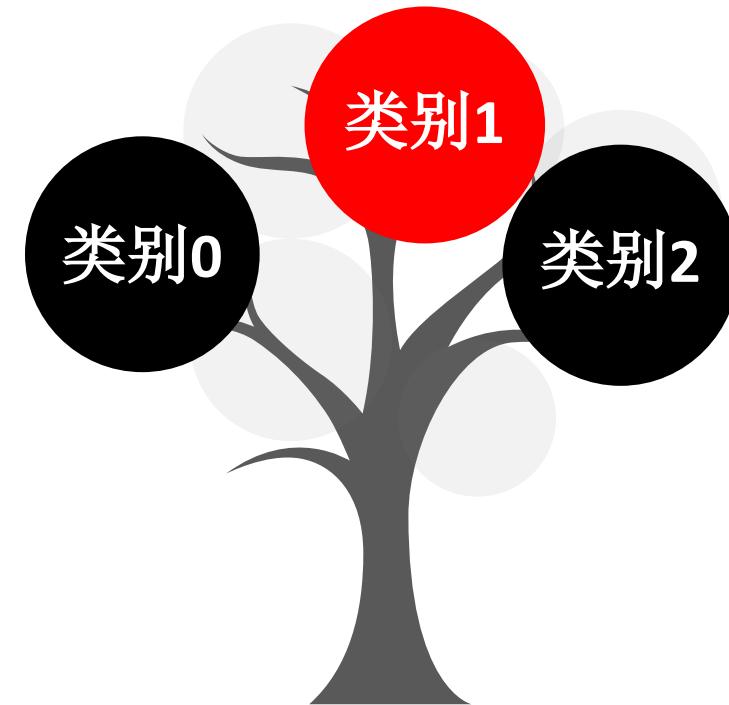
预测公式:  $Y = -34247X_1 + 331498X_2 + 39450X_3 + 710124X_4 + 499358X_5 + 348659X_6 - 395252X_7$

04

# 聚类模型以及评论分析

# K-Means聚类分析 (将1200个视频聚成3类)

预测分类	0	1	2
综合得分	4.707347e+05	133857.882040	2.356752e+05
点赞数	7.887094e+04	11662.557917	5.069292e+05
投币数	6.904393e+04	7870.310308	6.759245e+05
收藏数	5.662843e+04	8387.326249	4.844095e+05
分享数	1.070235e+04	1045.618491	1.518757e+05
评论数	4.072193e+03	766.615303	3.161043e+04
播放量	1.503722e+06	239549.580234	1.217264e+07
弹幕数	1.354827e+05	15895.443146	1.182756e+05
类别排名	3.397131e+01	54.716259	5.250000e+01
全站rank	9.968852e+01	986.181722	7.614286e+01



\*类别0：综合得分最高，类别排名最高

\*类别1：表现平平，各类指标均远低于另外两类

\*类别2：播放量最高，粉丝互动量最高



# 评论信息的情感分析模型

```
1 def emotion(s):
2     positive=0
3     negative=0
4     smooth=0
5     for i in s:
6         if i>0.6:
7             positive+=1
8         elif i<0.4:
9             negative+=1
10        else:
11            smooth+=1
12    counts=positive+negative+smooth
13    print('积极情绪: ', str(int(positive/counts*100))+'%')
14    print('消极情绪: ', str(int(negative/counts*100))+'%')
15    print('平和情绪: ', str(int(smooth/counts*100))+'%)
```

\*大于0.6的为积极情绪  
0.4-0.6的为平和情绪  
小于0.4的为消极情绪

```
1 len(text)
```

324795

```
1 a=[np.round(SnowNLP(i).sentiments, 2) for i in text[:5000]]
```

```
1 emotion(a)
```

积极情绪: 41%  
消极情绪: 18%  
平和情绪: 39%

- ◆ 从三万条评论信息中取出五千条样本通过SnowNLP进行情感分析，输出结果显示B站评论信息观众情感倾向依旧是以积极情绪为主，消极情绪仅占18%
- ◆ 后续通过对12种类别的视频评论进行分析，所得结果也大致与总体样本一致，积极情绪比例依旧最高为40%左右，消极情绪最低为18%左右
- ◆ 说明B站评论人群大部分是对所观看视频有正向态度或中立态度



05

总结

# 结 论

## 数据统计可视化

1

从发布数量和发布年份来看，高质量视频或集中于2014-2016年，数量激增始于2018年，如何维持在美上市后数量和质量并行是B站面临的挑战。在统计的十余种类别当中，鬼畜视频占据B站各项头牌，以此为B站带来巨大流量，增加并稳定鬼畜类视频发布数量与质量，是维持B站持续兴盛的方式，普及舞蹈、国创等小众视频的宣传与大众关注度将成为B站的一大挑战，应从“头重脚轻”的发展方式向更加平衡的发展过度，并持续把握新的热点

2

基于主成分分析的多元回归结果表明，弹幕数过高会影响视频的播放量，而投币数，点赞数，分享数，收藏数等维度均对播放量有正向的影响，提取主成分后的回归公式表明收藏数和投币数对视频播放量贡献度最高为0.46，综合得分对播放量影响最大为-0.84，综合得分越高的视频播放量反而越低，此维度数据会对用户选择质量高的视频造成干扰

3

## 评论分析

评论数据大多为积极情绪，说明B站作为年轻人的平台传递的情绪大多数为正向积极的态度



# 感谢您的收看

汇报：王储

时间：2019.09

