B站分区视频数据分析

分析背景

B站作为二次元网站，是年轻人的喜爱，作为小up主，如果想要**转型或者得到一些播放认可**，需要研究B站用户对哪些类型的作品比较喜欢，又或者另辟蹊径，在小众区独领。

翻看B站分区的TID，不难发现很多的单独的分区都已经消失了，由于分区活跃人数太少，分区也越变越少，逐渐被并入其他分区，不配拥有姓名。

数据来源

爬虫各分区**周排行前十**视频的数据，执行定时任务爬取一个月视频算其。

H5 分区情况

B站分为十几个大分区，大分区下若干分子区。

https://github.com/SocialSisterYi/bilibili-API-collect/blob/master/docs/video/video_zone.md

H5 爬取代码

mytest > Bilibilistudy > bilibiliscr.py > ...

```python
import requests
import  json
import time

data={}

with open("constants.json",encoding='utf8') as f:
    regions = json.load(f)

for rid,regions in regions.items():
    print(rid,regions)
    #resp = requests.get(f"https://https://api.bilibili.com/x/web-interface/ranking/v2?rid={rid}") 热门排名
    resp = requests.get(f"https://api.bilibili.com/x/web-interface/ranking/region?day=7&original=0&rid={rid}")
    d = resp.json()
    if d['code'] == 0:
        data[rid] = {"region" : regions, "data":d['data']}
    time.sleep(0.5)

with open('data.json','w',encoding='utf8') as f:
    json.dump(data,f)
```

H5 数据格式

'分区', '播放数', '投币数', '收藏数', '评论数', '时长'

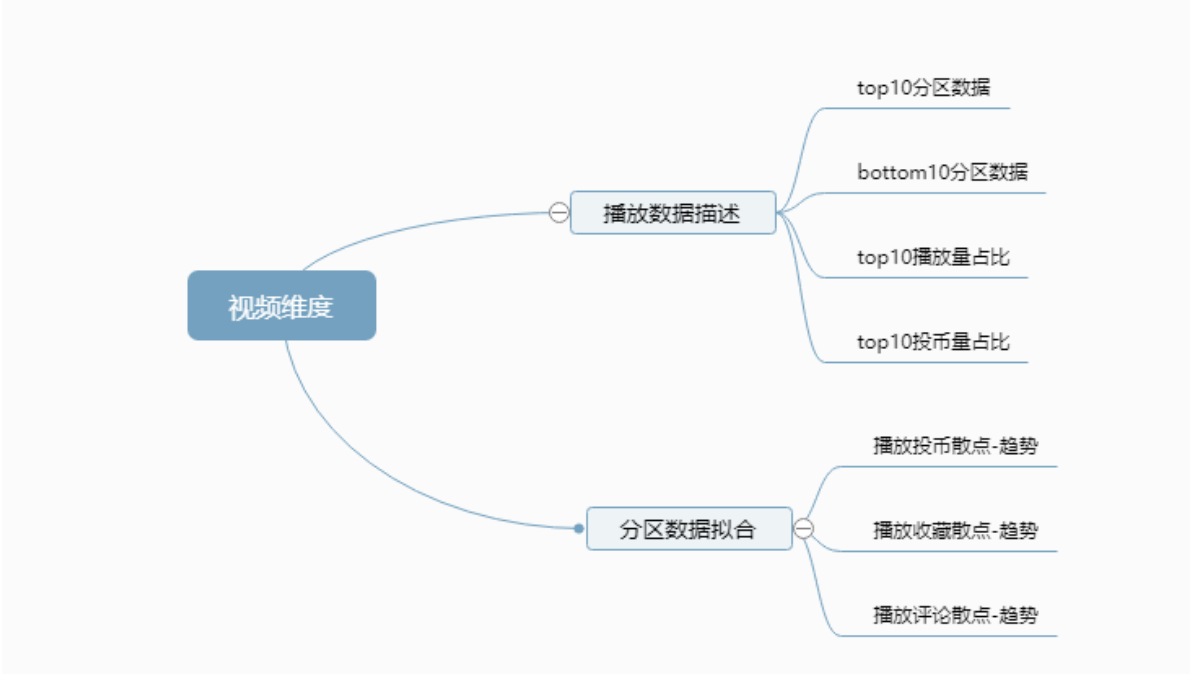
可能的问题

哪些分区观看流量大？为什么会这种情况？ 关键是什么？

由于视频**计算播放的规则**是：一个账号一天之内，只能给每个视频加一播放量，且要观看到视频的一半才行，是否会趋于短视频时代？

分析思路

数据来源仅有分区视频的播放投币等直观数据，从该方面出发。



数据处理

json格式数据的清洗，写入csv。

mytest > Bilibilistudy > data_ana.py > ...

```python
import  json
import csv

data = {}

with open("data.json") as f:
    data = json.load(f)

with open("data_csv.csv","w",encoding='utf8') as csvfile:
    data_csv = csv.writer(csvfile)
    data_csv.writerow(['分区','播放数','投币数','收藏数','评论数','时长'])
    for d in data.values():
        if all(d['region'] in item['typename'] for item in d['data']):
            play = sum(item['play'] for item in d['data'])
            coins = sum(item['coins'] for item in d['data'])
            favorites = sum(item['favorites'] for item in d['data'])
            review = sum(item['review'] for item in d['data'])
            duration = sum(int(item['duration'].split(':')[0]) for item in d['data']) + sum(int(item['duration'].split(':')[1]) fo
            data_csv.writerow([d['region'],play,coins,favorites,review,duration])
```

- 通过jupyter 总览数据情况，确保数据完整性，由于是json格式数据，基本是完美的。从这里看到B站总共有106个小分区。

```
In [29]: df.info()
```

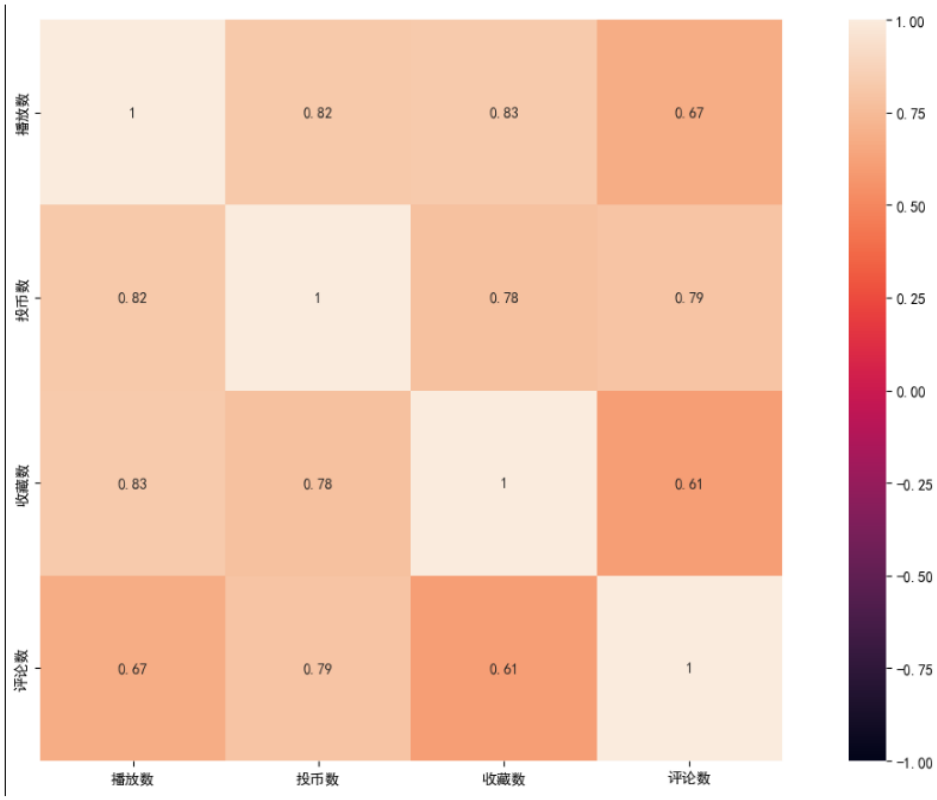
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 106 entries, 0 to 105
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   分区      106 non-null    object
 1   播放数     106 non-null    int64 
 2   投币数     106 non-null    int64 
 3   收藏数     106 non-null    int64 
 4   评论数     106 non-null    int64 
 5   时长      106 non-null    int64 
dtypes: int64(5), object(1)
memory usage: 5.1+ KB
```

数据可视化

1. 相关性分析

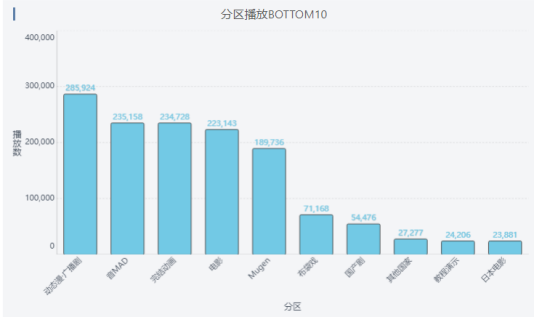
```
In [34]: train_corr = df.drop('时长',axis=1).corr(numeric_only = True)
print(train_corr)
a = plt.subplots(figsize = (16,9))
a = sns.heatmap(train_corr, vmin=-1, vmax=1 , annot=True , square=True)
```

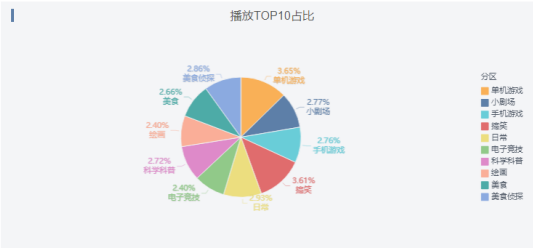
```
          播放数       投币数       收藏数       评论数
播放数  1.000000  0.820124  0.825009  0.672375
投币数  0.820124  1.000000  0.776951  0.789862
收藏数  0.825009  0.776951  1.000000  0.606570
评论数  0.672375  0.789862  0.606570  1.000000
```

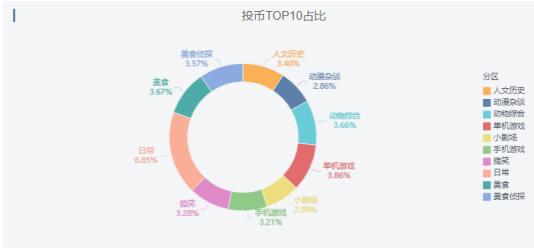


由于播放量意味着视频被更多人看到，所以几乎与投币数，收藏数，评论数是正相关的，切其他两两之间也是如此。总的方向如此，后续展开对不同分区之间的差别分析。

2. TOP，BOTTOM数据描述及其占比



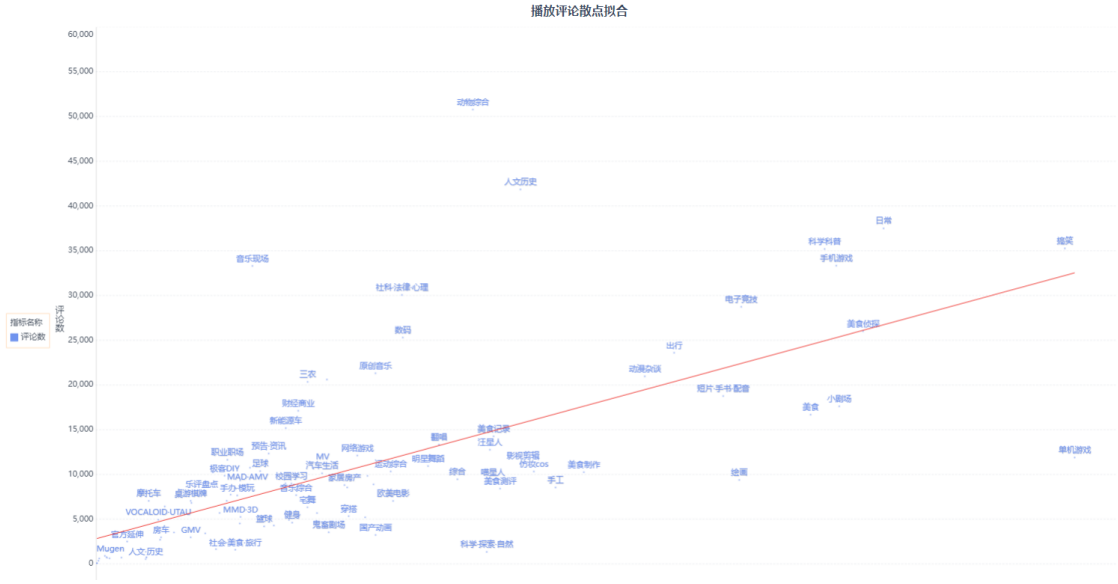




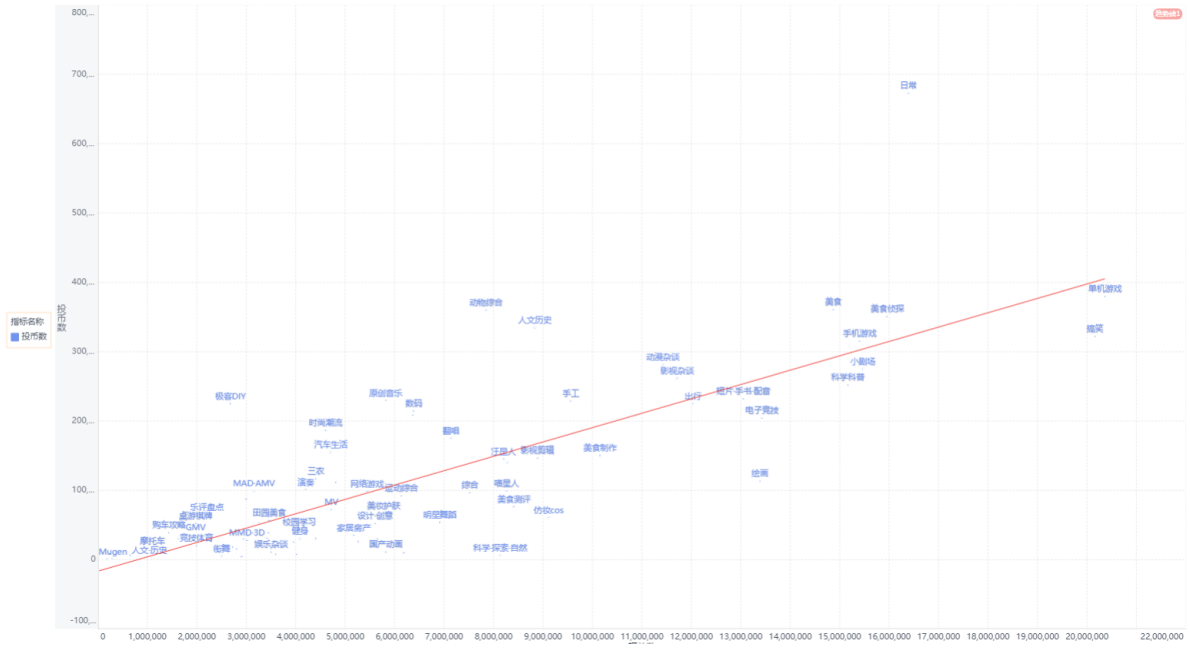
TOP，BOTTOM数据展示出，**游戏，日常，美食，绘画**等等播放量排名靠前，**电影，电视剧，教程类**的不受欢迎，也很反映B站ACG的特色，二次元浓度比较深。

在播放TOP和投币TOP中，抛开播放量对投币的正相关性，**日常生活类和人文历史类**脱颖而出，说明B站用户对这两大块是很认可的。

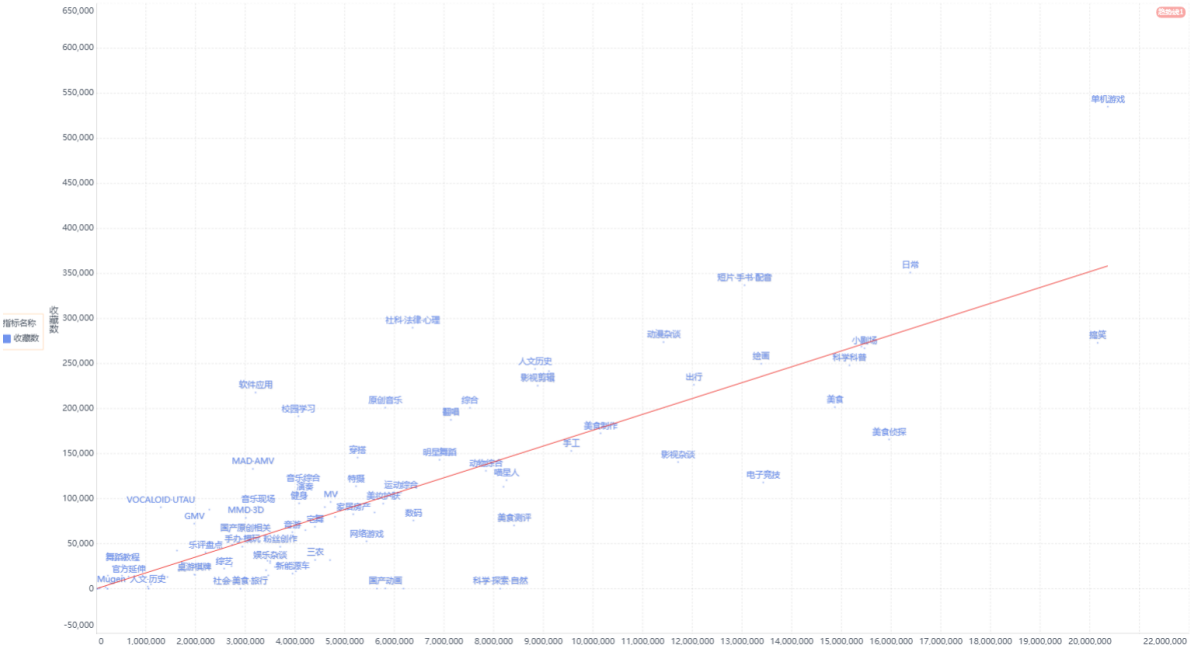
3. 播放评论收藏拟合



播放投币散点拟合



播放收藏散点拟合



散点拟合解析：拟合线代表了平均绝对值

播放数：其受众人群，大众---小众

投币数：认可度　高---低

收藏数：价值量　高---低

评论数：话题性　高---低

所以三张图都可以用一种思维去解读：

在拟合线左上的，是小众的高**认可度\价值量\话题性**视频，但这类视频有潜力，很值得发掘。
右下的是大众的，低**认可度\价值量\话题性**视频，但这类视频有趣，是生活的调味品。

用人的爱好、圈子文化去理解，圈子越小其受众人群越小，不仅有门槛，还往往有极强的凝聚力，而生活类，搞笑类等等受到推广，普通人也能观看。

投币数拟合图：日常，历史，极客DIY，动物综合，原创音乐，数码是属于观众认可高的。在这其中，极客DIY，数码，原创音乐更是小众的高认可，是值得人深挖的；而国产动画，妆仿则是较低的认可度。

收藏拟合图：单机游戏，短片手书配音，社科法律心理，软件应用，校园学习等等，太多的收藏分区涉及教育学科类学习的方法，资料，这类有着极高的价值，但除开单机游戏，播放量都很低；而搞笑，美食，网络游戏，电子竞技等等却是右下角的消遣类，不含太高价值。

评论拟合图：人文历史，音乐现场，动物综合，社科心理法律是容易引起人们互动的分区。

还有科学自然区是仅有播放，其他数据都不怎么样的分区。

B站平台有它的视频算法推广，导致**“强者俞强”**的情况，所以才会有“号养好了，打开推荐都是”相关的言论；还有热门视频多，而硬币一天只有一个，无法投给多个视频的情况；一键三连的选项，也会导致收藏一些没有太多价值的视频等等。所以，需要综合三个指标去衡量视频的优秀与否，如何分配权重就没有展开了。

打开数据来源的GITHUB，会发现已经有不少分区被划掉，属于它的TID也永远消失了。小众的有门槛的分区得不到热度，分区顶梁up出于爱好的视频评论下永远是老用户，那些6级用户，短位ID，视频得不到推广，那么也就不会有新人点进去入坑，逐渐消失。大众的经过算法推广越来越多人沉浸其中，乐此不疲。