# Bilibili城市视频统计数据分析

在之前的城市数据挖掘中，总共挖掘了294个城市的总共5880个视频的弹幕信息和基本播放数据。

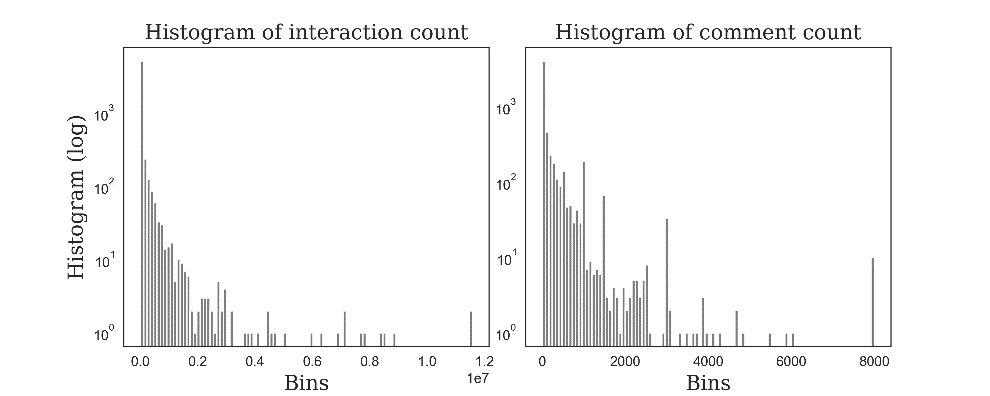
1. 数据清理

虽然本作业中爬取的弹幕数据来自官方的网站，大部分格式均是符合要求的格式，只有少数额外的弹幕包含了不是弹幕文本相关的信息，如额外的格式符号或者特殊函数之类的，如下图的两个例子：

* '["0.107","0.217","1-0.98","10","新宝岛！！！","0","5","0.99","0.654","1800",0,1,"SimHei",0]'
* 'def text v {\n content = "233"\n fontSize = 5%\n fontFamily = "SimHei, Microsoft JhengHei"\n bold = 0\n textShadow = 1\n strokeWidth = 0\n strokeColor = 0xFFD302\n zIndex = 0\n anchorX = 0.5\n anchorY = 0.5\n x = 50%\n y = 50%\n color = 0xFFFFFF\n scale = 1\n alpha = 1\n rotateX = 50\n rotateY = 30\n rotateZ = 30\n}\n\nset v {\n} 3s\n// Generated by bas visual panel'

因此为了减少排除这些影响长度估计的弹幕，设计了正则表达式将类似的弹幕去掉了。

1. 总体视频数据分析



首先从总体上分析视频数据，从直方图观察视频播放数量和弹幕数量的分布情况。从上图看，大部分的播放数量在以内，弹幕数量则在内。从结果上看，分布基本属于一个常规的长尾分布的样式。

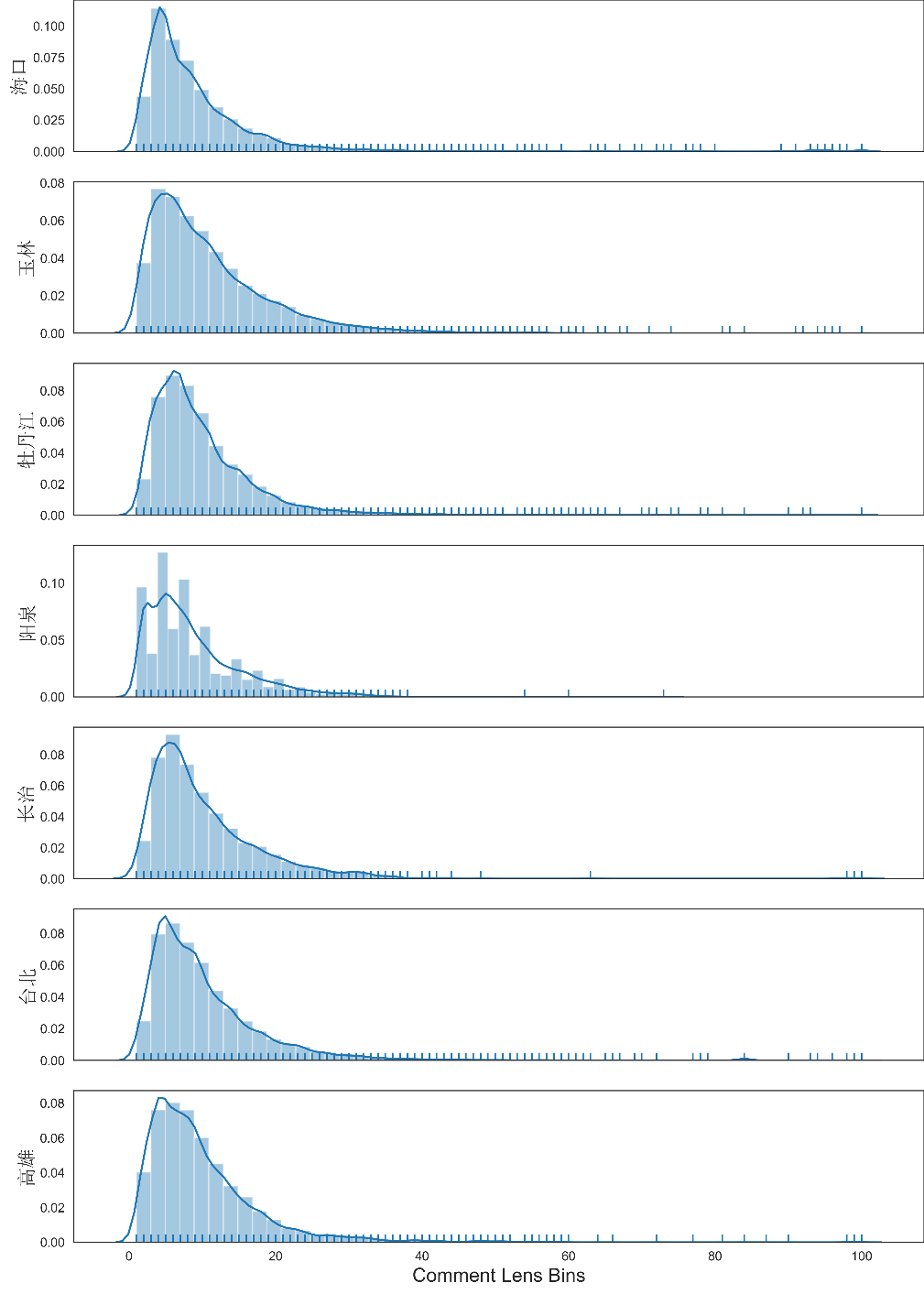
从播放数量直方图中可以看到最右侧有两个很高播放量的视频，以及弹幕数量最高的视频，在数据中找出其信息如下表所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Video Url | City | Video Name | Interaction  Count | Comment  Count |
| 播放量最高 | https://www.bilibili.com/video/av61089242 | 中山 | 不要惹穿中山装的男人！ | 11576622 | 1000 |
| https://www.bilibili.com/video/av12236018 | 海口 | 海口蹦迪王 | 11583500 | 300 |
| 弹幕量最高 | https://www.bilibili.com/video/av38189520 | 玉林 | 东方直播室：舌尖上的风波 (关于玉林荔枝狗肉节一期全) | 186036 | 8000 |
| https://www.bilibili.com/video/av880349 | 牡丹江 | 爸爸去哪儿 20131220 牡丹江雪乡特辑上 | 378058 | 8000 |
| https://www.bilibili.com/video/av889114 | 牡丹江 | 爸爸去哪儿 20131227 牡丹江雪乡特辑下 | 293178 | 8000 |
| https://www.bilibili.com/video/av4819260 | 阳泉 | 人说山西好风光 市长真人秀 阳泉 朔州 临汾 160527 | 51892 | 7993 |
| https://www.bilibili.com/video/av4856678 | 长治 | 人说山西好风光 市长真人秀 吕梁 太原 长治 大同 160603 | 109656 | 7997 |
| https://www.bilibili.com/video/av2979957 | 台北 | S.H.E. 2gether.4ever.台北安可场（2014） | 334963 | 8000 |
| https://www.bilibili.com/video/av6227667 | 高雄 | 2015林俊杰「时线•新地球」世界巡回演唱会-最终站高雄站 | 184751 | 7989 |

从视频的名字可以看出，部分视频属于搞笑类的视频，部分视频是综艺节目，还有部分是一些演出的视频。而从他们关联的城市上看，除了中山不是城市相关视频外，剩下的海口、玉林、牡丹江、阳泉、长治、台北、高雄均为城市。可以发现，台湾的城市因为举办了一些演唱会而出名，山西省则是政府进行了宣传，牡丹江则是因综艺节目而收到关注。

1. 样例城市弹幕长度分析

从上述的分析中了解到海口、玉林、牡丹江、阳泉、长治、台北、高雄这些城市拥有很多弹幕，而关于弹幕的统计信息比较局限，这里以它们的长度来看弹幕长度的分布情况，如下图所示。



根据上述的长度分布可以看出它们均符合一致的一个类似高斯的分布，说明了弹幕分布的一个广泛规律：均值为9.91，标准差为 8.21。