Final PJ: 黑网吧可视分析

数据清洗：

（1）删除网吧名称中的括号和空格

（2）将出生日期在 2017.1.1 之后或出生日期格式出错的上网记录删除

（3）删除地区 id 格式有误的上网记录

（4）将上线时间大于下线时间的上网记录删除

问题 1：请对某市网吧上网记录进行分析，从中发现非法经营现象（接纳未成年人上网）。由于接纳未成年人需要使用成年人有效证件帮助其进行实名上网登记，试着找出用于接纳未成年人的成年人信息，并展示和说明未成年人上网接纳情况。

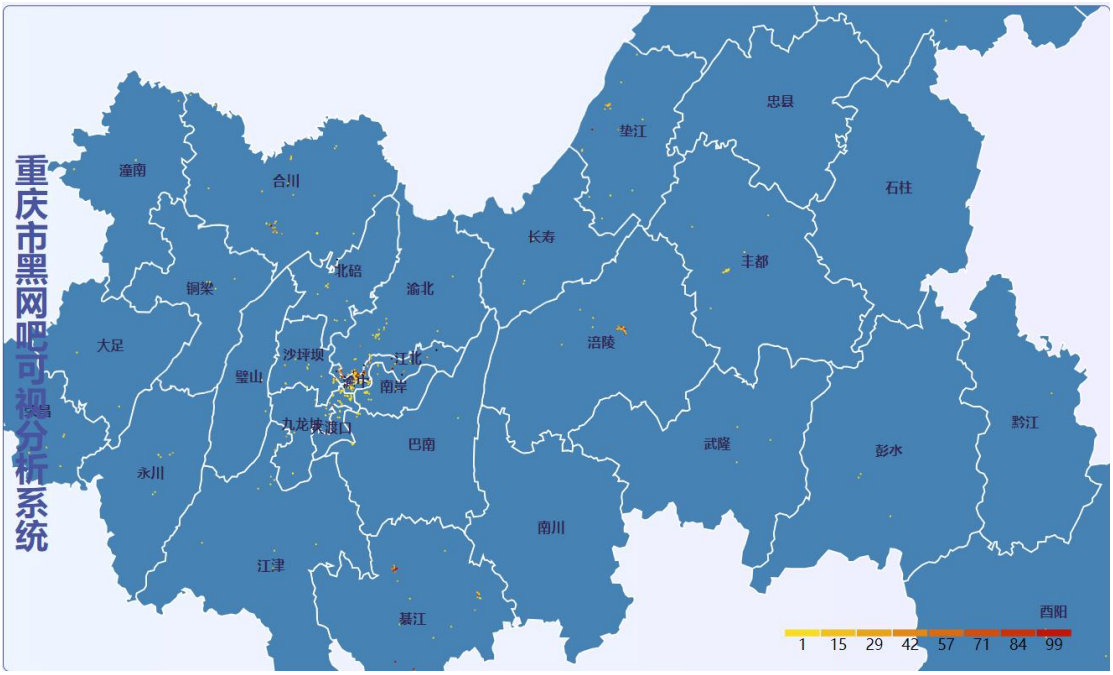
（1）将连续在线时长多于 168 小时（7 天）的成年人视作被用于接纳未成年人的成年人信息，部分成年人信息如下：


类别: 成年人

ID	姓名	违规次数
4fd1623fbda4f9ad5a	李**	2
b8d6e4e3d216b79161	王**	1
3f6c45b4b42889d0b9	谢**	1
2f7c5b8035b0bfc6f1	李**	1
e5a2015e1e96b96a2d	马**	1
00321c8ed21c582ac1	汪**	1
41bc4a2011d1094c86	王**	1
9d08bdcc62a3e7a136	王**	1
50ce5dea7862eb3843	梅**	1
41802e54ea619479b5	任**	1

（2）

（a）网吧违规接纳情况如下：



其中颜色代表违规接纳的次数，违规次数随着颜色由黄到深红增加。可以看到，违规的网吧基本集中在渝中，江北，南岸，沙坪坝地区，这是因为这几个区靠近市中心，网吧数量多；同时万州、涪陵、綦江和合江区也有较多的违规网吧，这里可能是交通枢纽和人员集散地。

（b）将鼠标悬停至网吧的位置（圆点）上，即可显示网吧的具体信息，如图所示：

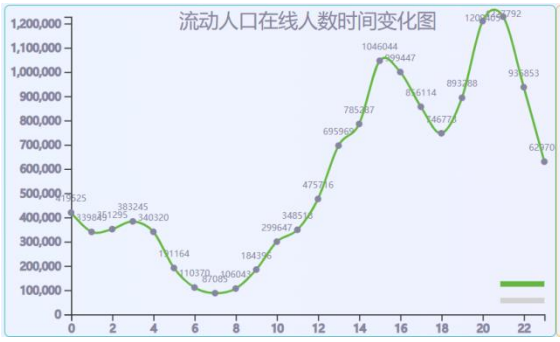
（c）

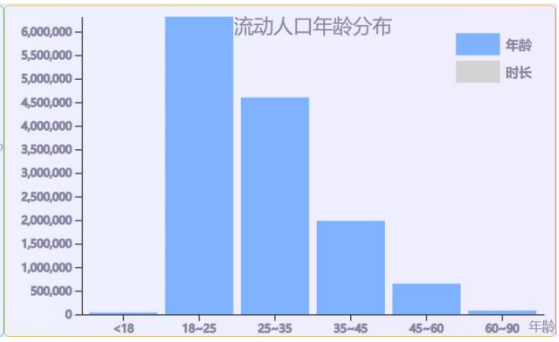
部分违规的未成年人信息如下：

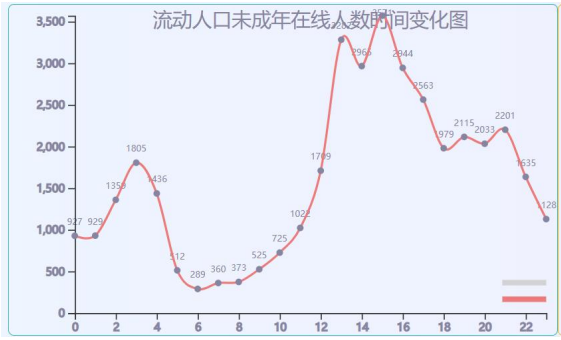
类别: 未成年人

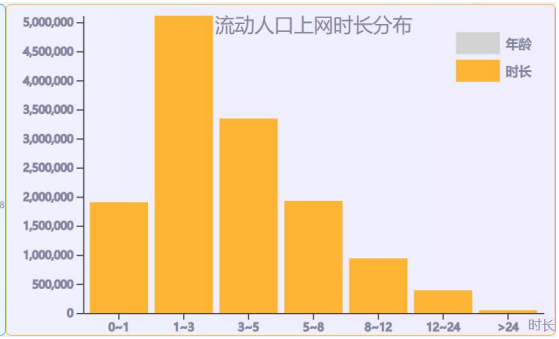
ID	姓名	违规次数
c45dcaaf87b5cf7692	余**	6
5a0586d17c82d3be9a	王**	5
fb5208d283219cf5ea	赵**	5
d460ca919af75ffb0b	周**	5
d6a6116147ca62c2dc	赵**	5
65cb9dbe028d4e542d	周**	5
a4a575cf3dd9c9a692	曾**	4
82ed8c6e68b4733d96	赵**	4
fd68a8188ba07743e1	杨**	4
b0e4c4f416cb7c329a	叶**	4

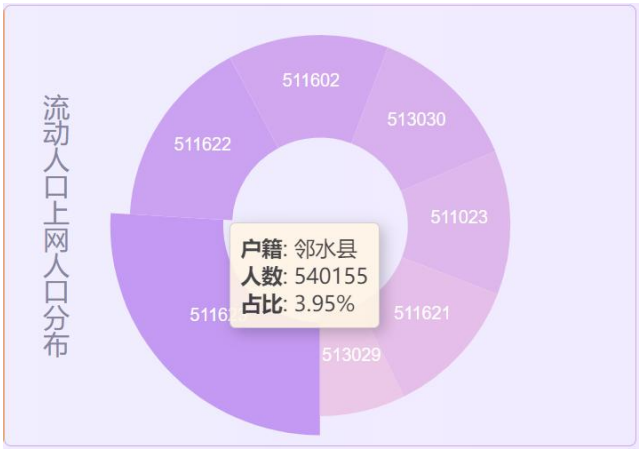
问题 2：流动人口（籍贯为非本市，题目中某市的籍贯代码前两位为 50）犯罪问题是我国工业化、城市化的伴生物。由于流动人口缺乏对非落户城市的归属感，容易因为心态不平衡而导致犯罪。试着分析流动人口的上网记录并总结他们的行为特点。











可以看到，流动人口上网的高峰期是在下午 14-16 点和晚上 20-24 点；年龄主要是 18-25 岁的年轻人；流动人口未成年人上网的高峰期主要是 13-15 点；上网时长主要在 1-3 小时；流动人口主要来自邻水县（3.95%）、武胜县（2.47%）、广安区（2.08%）。

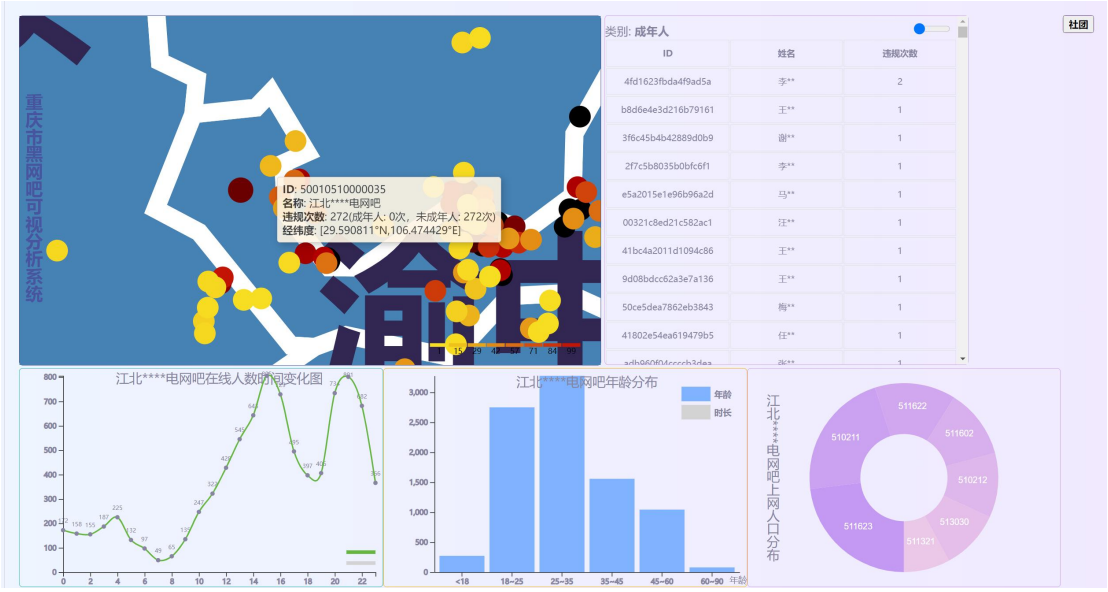
问题 3：青年犯罪团伙倾向于聚集在娱乐场所内，而网吧是唯一需要登记的娱乐场所。通过上网时空关系能够推断用户之间可能存在的联系，并辅助公安人员刑侦以及犯罪预防等工作。请试着从上网记录中发现社团。

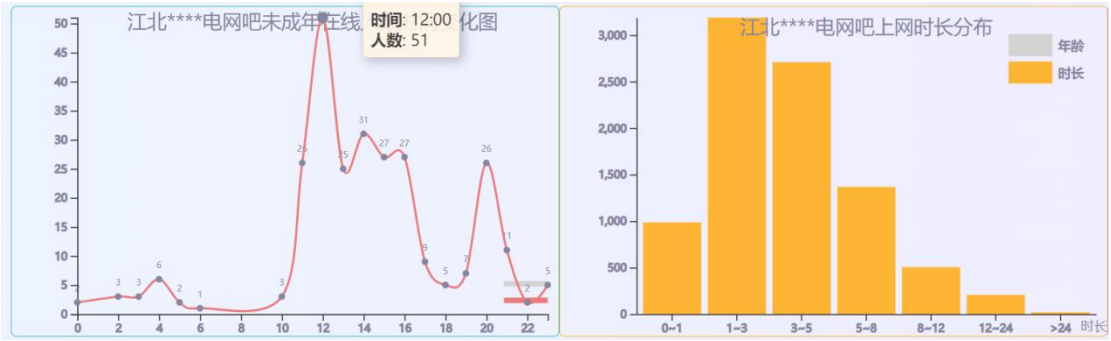
这里采取了滑动窗口的方式来寻找用户之间的关系：先找出上网次数大于等于 5 次的用户的上网记录，将这些记录按上线时间进行排序。将上线时间相差小于等于 15 分钟当作是相关，若在同一个网吧或下线时间相差也小于等于 15 分钟，则在两个用户之间增加权重 1，并以节点、权重的信息作如上的可视化图。

人数大于等于三的社团有：

1. f288b3a6f813c2cec9、0c20cb9c893efc39f0、56066dfa7fdbc8e17d
2. ba2f6e4d0f878b599b、84bdbbc7d421310fc0、42bad9ddc6b2d9c816、3e233da1a5c290bf94、080e48f900bd65ddb9、15e863e224b6a23092、74c42dc1fa15ae2bec、3183344d09c5a509b1
3. 6bc69f1aa2873c2054、 2009fc62a2777356f4、1286235e4929339e8a
4. 95b4bd8924cfb47fec、1ff1233c0db2a559ac、5596b31d57784cb5ba

问题 4：为了设计出目标人群喜欢的产品，产品经理常通过问卷调查、访谈和统计等方式，获得可以区分出目标人群的用户特征或者说用户画像。借鉴上述做法，公安人员可以为网吧做用户画像，可用特征有很多，比如：未成年人上网高峰时段、上网人群年龄以及外来人口比例等等。（可从上网时间、时长、上网人员籍贯等维度分析）请综合上面 3 个问题的分析结果，从多角度设计并展示网吧的用户画像。并且根据你所搜集的信息以及分析的结果，试着对你所在的公安局提出综合性建议。





以江北****网吧为例，可以看到该网吧用户上网高峰期和流动人口上网高峰期类似；年龄主要在 25-35 岁；未成年人在线高峰时段为 12 时和 20 时左右，这可能是因为未成年人在这一时间段放学；上网时长主要在 1-5 小时范围内；上网人员的籍贯中，占比较大的为邻水县（5.85%），江北区（5.57%），武胜县（3.52%）。

该网吧违规记录中都是直接接纳未成年人上网，因此公安局可在中午 12 时或晚上 20 时对网吧进行检查，并对违规行为进行处罚。