.

2017 年第四届中国可视化与可视分析大会

数据可视分析挑战赛-挑战 2

（ChinaVis Data Challenge 2017 - mini challenge 2）

答　卷

参赛队名称：东北大学-高强-挑战 2

团队成员：高强，东北大学，1045010454@qq.com，队长

黄健，东北大学，435219253@qq.com

周欣娜，东北大学，1020653488@qq.com

刘昕蕊，东北大学，461593398@qq.com

姚羽，东北大学，yaoyu@mail.neu.edu.cn，指导老师

是否学生队（是或否）：是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）:Gephi，Echarts，MySQL，MyEclipse，Excel，百度地图

共计耗费时间（人天）： 40 人天

本次比赛结束后，我们是否可以在网络上公布该答卷与视频（是或否）：是

（灰色字为参赛信息填写模板，请参赛者在提交时参照模板填写）

挑战 2.1：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。（建议参赛者回答此题文字不多于 1000 字，图片不多于 8 张，可使用附录形式列出非法网吧和非法使用的成年人信息）

根据所给数据，我们将分析后的全部信息可视化展示在如图 1.1 所示的网页中，集成为一个重庆市网吧的可视化分析系统。

图 1.1 重庆市网吧可视化分析系统

同时，系统也具有一定的交互能力，点击某网吧，如图 1.2 所示，将会展示其详细信息。

图 1.1 某网吧详细信息

非法经营网吧大体可分为直接接纳未成年人和非法利用其他成年人的身份证号帮助未成年人上网两种情况。

其中，我们把有非法经营现象（接纳未成年人上网）的网吧展示在地图上，如下图 1.3 所示，蓝色代表直接接纳未成年人上网的网吧，黄色代表违规利用某些成年人身份证号帮助未成年人上网的网吧，点的大小代表违规上网人数的多少。

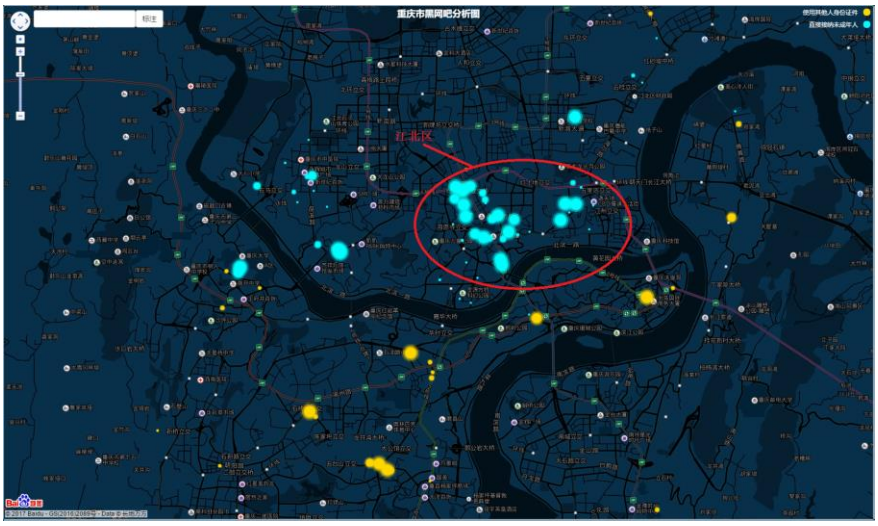


图 1.3 显示在地图上的违规网吧

具体的分析过程与结果如下：

①直接接纳未成年人

有一些网吧无视法律法规，直接接纳未成年人上网，在地图上以蓝色圆点表示，点击某一网吧，可查看其详细数据。

如图 1.4 所示，横坐标代表时刻，纵坐标代表上网人次，红色折线代表未成年人，浅蓝色折线代表年龄介于 18 和 60 岁之间的人群。通过下图可以看出，这家位于江北区 id 为 50010510000039 的江北****会所，在三个月内的 16 时总共接纳了未成年 615 人次，是一家毫不掩饰的“黑网吧”。

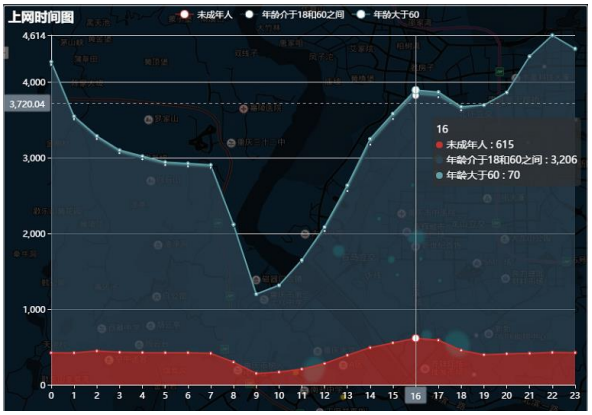


图 1.4 某网吧的上网时间图

②根据年龄进行分析

我们设定 40-60 岁年龄段且连续上网时间大于 72 小时或 60 岁以上连续上网时间大于 72 小时的用户为“被套牌”用户，考虑到这些年龄段人群的身体状况、生活背景和普遍兴趣爱好等因素，我们有理由怀疑这些上网记录所在的网吧是利用这些身份信息接纳未成年人上网的非法经营网吧。

如图 1.5 所示为部分年龄大于 60 且单次连续上网时长大于 72 小时的用户，我们用列表的方式将其直观地可视化展示。例如，身份 ID 为 d1d3622d445fccd58a 的 88 岁张**先生,显然不能单次连续上网 149.25 小时。很显然，这是一个被违规利用的身份信息，我们有充分的理由怀疑接纳这个用户的网吧是一家“黑网吧”。

身份ID	姓名	年龄	单次上网最长(h)	上网总时长(h)	上网次数
d1d3622d445fccd58a	张**	88	149.25	150.75	2
d1d3622d445fccd58a	张**	88	149.25	150.75	2
ff53227ee59e36709d	郝**	88	146.01	146.01	1
0eae781ca8cda2179d	袁**	88	138.53	138.53	1
a1a58b484ea60bd89f	杨**	88	138.01	141.03	2
a1a58b484ea60bd89f	杨**	88	138.01	141.03	2
b619276f4fb78d358f	刘**	88	138.01	138.01	1
759259d519a1499a0c	韩**	88	101.05	101.05	1
cdae9d686e409fd308	黄**	88	101.05	101.05	1
0e7a265a569eea413b	况**	88	97	97	1
47239c229c833a732f	闫**	88	97	103.61	2

图 1.5 部分年龄大于 60 的“被套牌”身份信息

③上网单次时长过久

有一些年龄段正常，但上网时长过久的记录也引起了我们的怀疑。我们设定连续上网时长大于 1000 小时的上网记录为被网吧违规利用的身份信息。

据新闻报道，“由于网吧下机不需要关机，网费上完后会自动关机。所以，有些小孩专门在网吧内等别人离开时，就立马上前记住其座位号，然后跑到吧台给座位号付费加时。这样，就可以用此前那个人的身份上网了。”

如图 1.6 所示，如身份 ID 为 82b6361ab67251bbeb 的 34 岁肖**男士，单次连续上网时间达到 3782.6 小时，这显然不符合现实生活的常理，我们有充足理由怀疑他的身份信息被如上文所说的新闻中的手段所利用了。

身份ID	姓名	年龄	单次上网最长(h)	上网总时长(h)	上网次数
82b6361ab67251bbeb	肖**	34	3782.6	3783.84	3
9cf586587765475d21	肖**	34	3782.6	3795.22	2
19f257ba373e5cdf88	肖**	29	3782.46	3782.46	1
7f3ee0d55b2a4e301b	张**	29	3782.46	3809.54	2
45af8d7343ea48f2f1	冯**	35	3519.37	3519.37	1
0c4ddd34fc76887a40	谢**	26	3332.53	3332.53	1
f96f017fc2701e3d42	许**	26	3332.53	3332.53	1
0f40523c5d295992de	黄**	23	3329.8	3333.26	2
422c5d11f2e6637244	张**	23	3329.8	3347.87	3
2468955d566ec15ca4	袁**	25	3329.71	3330.16	2

图 1.6 部分上网单次时长过久的信息

分析非法网吧的维度有多种，但由于所给数据的限制，我们只挑选了以上三个方面进行分析，得到的非法网吧和非法使用的成年人信息请详见**附录 2**。

挑战 2.2：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（可从上网时间、时长、上网人员籍贯等维度分析）。（建议参赛者回答此题文字不多于 1000 字，图片不多于 8 张）

发现流动人口的核心是分析每个上网用户的区域代码，找到相应城市，我们应该明确重庆市在 1998 年改为直辖市之前的身份证区域代码为 5102，1998 年之后才更改为 50，所以由于各种历史原因，数据库中并没有籍贯代码前两位是 50 的数据。

在明确这些先决条件之后，我们再分析各个省市在重庆市的流动人口数量，并分析其行为。

首先，分析各个城市在重庆市的流动人口数量。如图 2.1 所示，每一条白色虚线都代表这个城市有人口流动到重庆市，由于数据量过大，我们选取 top12 的城市进行展示。

由下图 2.1 可以看出，由于重庆市是西南地区的中心城市之一，作为四川省曾经的重要城市，重庆市吸引着大量周边外来人口。重庆市外来人口由高到低的 top12 分别是：广安市、内江市、南充市、遂宁市、泸州市、宜宾市、成都市、绵阳市、广元市、自贡市、铜仁地区和毕节地区。

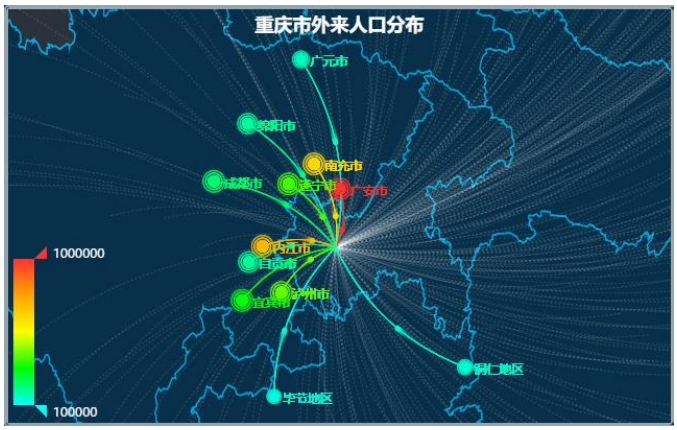


图 2.1 重庆是外来人口城市 top12

其次，我们对重庆流动人口的籍贯省份进行分析，利用条形图进行展示。橙色代表女性，蓝色代表男性。如下图 2.2 所示，重庆市外来人口的省份集中在周边省份，top11 的省份分别为四川省、贵州省、湖北省、湖南省、河南省、江西省、云南省、安徽省、陕西省、江苏省和甘肃省。这与重庆市“东邻湖北、湖南，南靠贵州，西接四川，北连陕西”的地理位置有着密切关系。

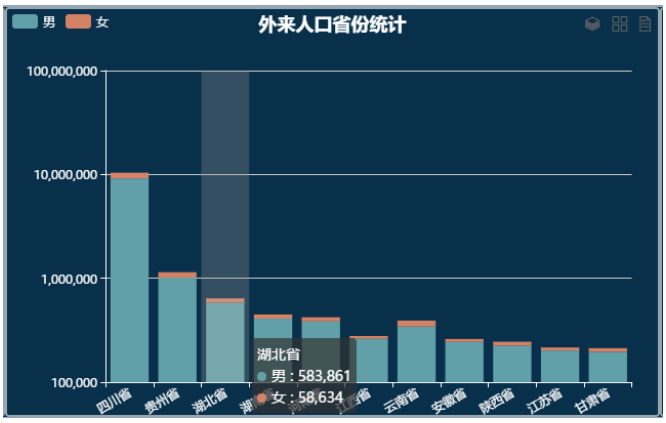
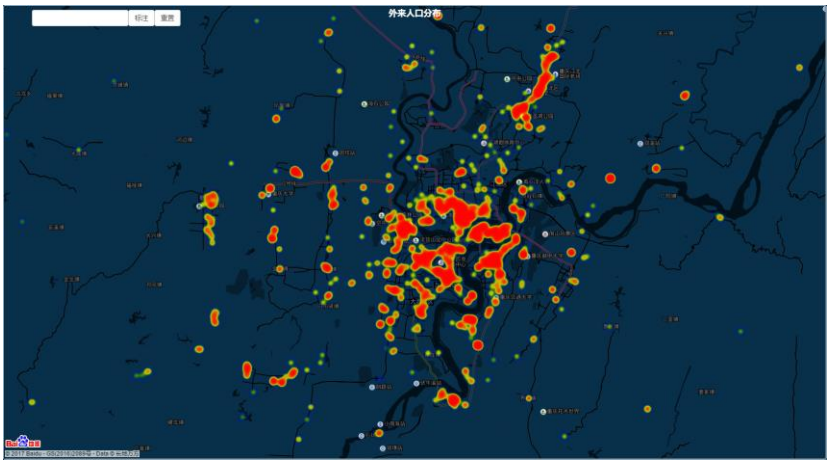


图 2.2 重庆市外来人口省份 top11

再次，分析各省市外来人口在重庆市活动的区域范围，热力图与 2.2 的条形图对应，下图 2.3 是重庆市外来人口的热力图，由绿色到红色代表人口活动愈加频繁。



再次，利用散点图展示外来人口在三个月内的上网的高峰时段。

如图 2.3 中的点代表的所有网吧的外来人口上网高峰时段，，图中点代表的是星期六晚 7 时的外来人口上网记录总条数为 109398 条。借此可以简要地分析出流动人口的上网高峰时间是每日中午的 12 时到晚上的 10 时，高峰日期是每周五、周六和周日。

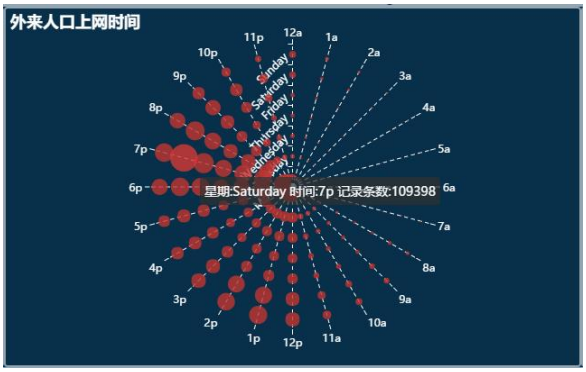


图 2.3 外来人口上网时间散点图

最后，为了分析外来人口活动的空间范围，在地图上点击某一网吧，显示如图 2.4 所示的扇形图，分析了该网吧各年龄层和流动人口的比例。

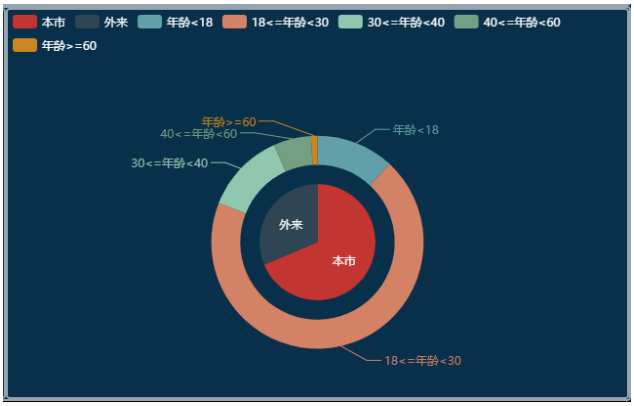


图 2.4 某网吧不同类型用户比例

挑战 2.3：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。（建议参赛者回答此题文字不多于 1000 字，图片不多于 8 张，可使用附录形式列出发现的社团）

在确认团伙的部分，我们采取了一种基于可变滑动窗口的数据挖掘方法。

首先，筛选出上网次数大于 5 的用户（所有数据中最大上网次数为 22 次），将这些数据按照上线时间升序排列。

其次，设限制窗口大小为 100，将窗口内第一条数据的下线时间与其他数据的下线时间进行对比，将上下线时间差均在正负 15 分钟内的数据赋予权重为 1，并进一步对比这些数据是否在同一家网吧内，如果是，权重+1，如果不是，权重保持不变。

最后，利用 Gephi 得到如下图 3.1 所示的拓扑图，点的大小代表上网次数，线段的粗细代表用户间关系见的强弱关系，黄色点代表发现的三个团伙。

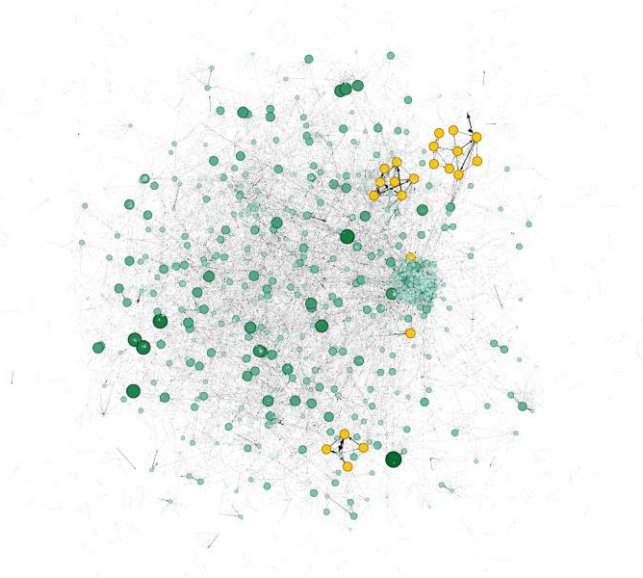


图 3.1 用户关系拓扑图

以在某一没有登记在册（单只数据库内不存在）网吧活动的某团伙为例，分析其年龄，籍贯均相同，可以初步确定他们为一个团伙，分析他们的行为，如下图 3.2 所示。

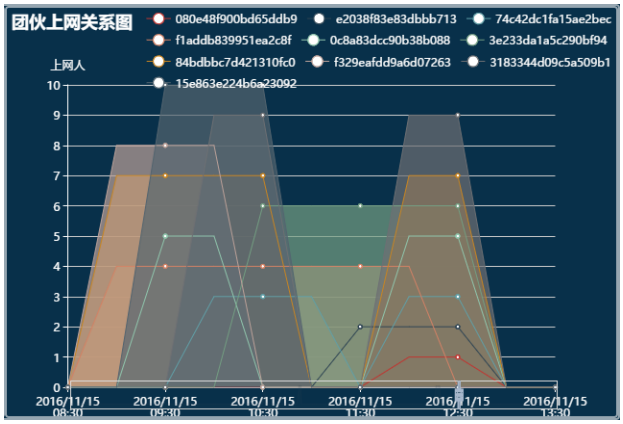


图 3.2 某青年团伙上网关系图 1

如图，纵坐标上的每一个数字代表团伙内的一个人，横坐标代表时间，每条折线代表一个团伙成员（用户 id）的上网过程。例如，2016 年 11 月 15 日早 9:30，0 号、4 号、5 号、7 号、8 号和 10 号 6 名团伙成员同时上线（间隔小于 15min），10:30 的时候 6 号成员和 9 号成员上线。重叠的部分可以有效且直观地显示出该团伙成员的聚集时间范围，对分析其团伙行为十分有益。

如图 3.3 所示，是该团伙在另一个时间段的上网关系图，可见该团伙在 2016 年 11 月 22 日的活动人员和上网时长都不如 2016 年 11 月 15 日频繁，但由于这两次都是周二，可以初步猜测该团伙的团体活动日期为周二。

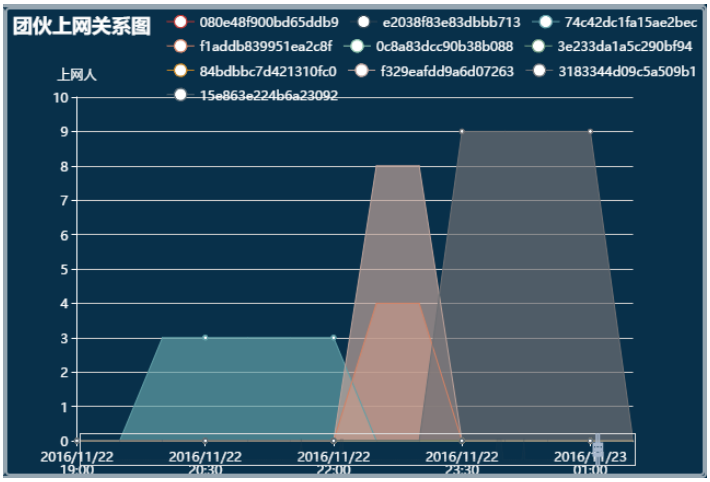


图 3.3 某青年团伙上网关系图 2

挑战 2.4：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。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。（建议参赛者回答此题文字不多于 1000 字，图片不多于 8 张）

综合以上 3 道题的结论，我们可以简要分析违规网吧所处的位置多分布于学校和商业街附近。如图 4.1 所示，蓝色点代表直接接纳未成年人上网的网吧，黄色点代表违规利用某些成年人身份证号帮助未成年人上网的网吧，红色点则是利用百度地图标记出的学校和商场。可以看出，位于市中心的商业区附近，违规网吧分布最为密集。

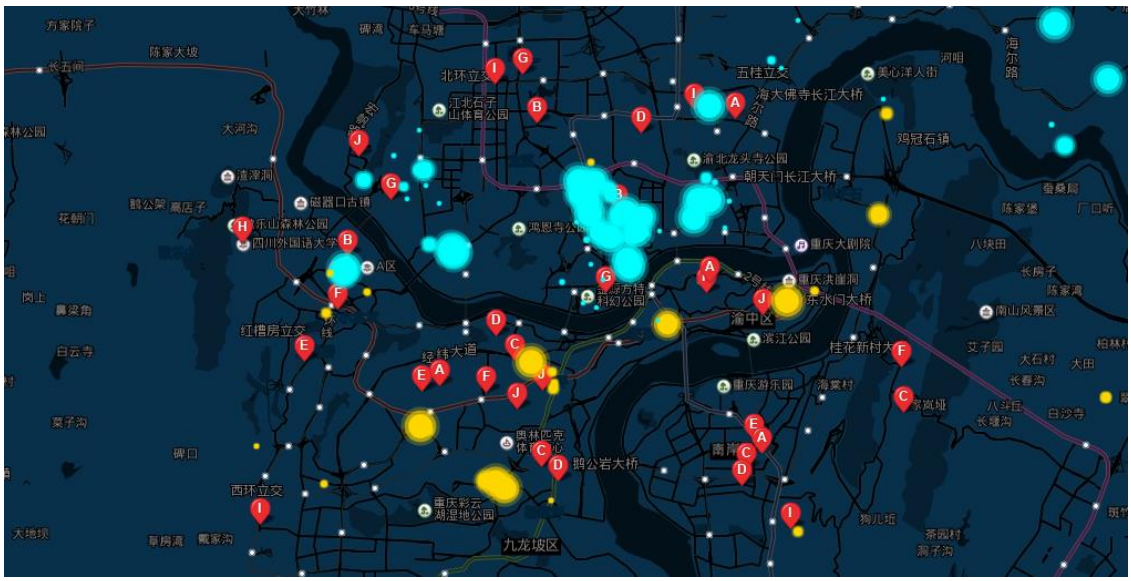


图 4.1 违规网吧和标记地点

我们继续详细地分析用户们的上网高峰时段，根据年龄进行划分。如图 4.2 所示，左图依然是位于江北区 id 为 50010510000039 的江北****会所吧，我们可以看出直接接纳的未成年人的高峰时段为下午 16 时之后。右图为位于郊区的 id 为 50011710000064 的重庆****联网吧，接纳未成年人的高峰时段为晚 20 时之后。

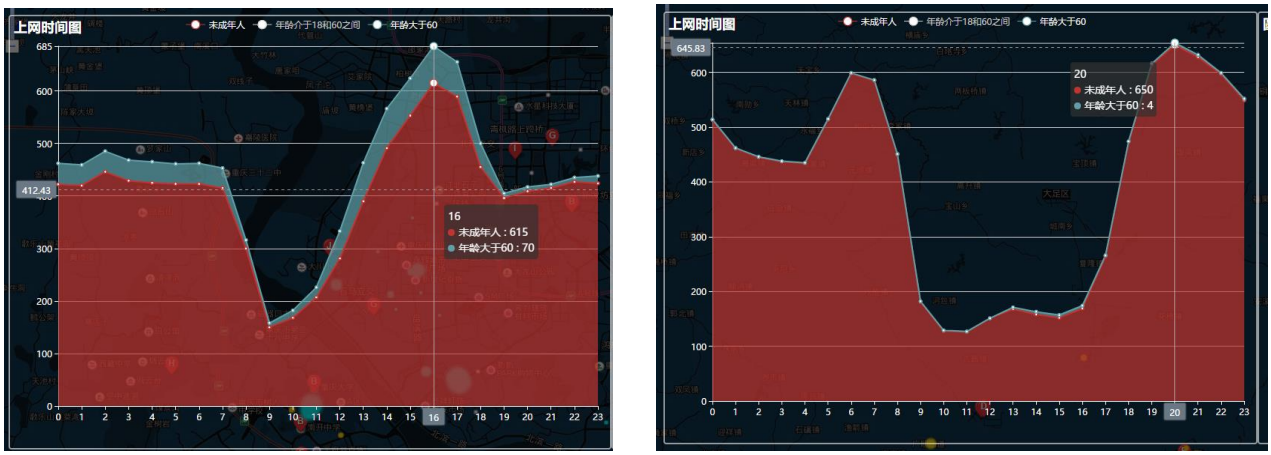


图 4.2 市区和郊区网吧高峰期对比

据此，我们可以分析，由于交通、学校分布、家庭住址等多方面原因，郊区网吧的每日高峰期到来时间要晚于城区内网吧的高峰期。

有关三个月内的日历高峰图，如图 4.3 所示，横坐标代表日期，纵坐标代表星期（第一行代表 Sunday 周日）。将三个月内的每日上网记录数量，按比例显示在图中，选取 Top 12 的点高亮显示。我们可以发现，由于 10 月份有较长的国庆假期，12 月份又临近年底，普遍较忙，上网的高峰期主要集中在 11 月份的普通周五和周末，这与我们的实际生活规律相符合。

图 4.3 上网记录日历图

除此之外，我们综合考虑外来人口的省份、上网次数、年龄和平均上网时长，做出了如图 4.4 所示的多维度动态散点图。

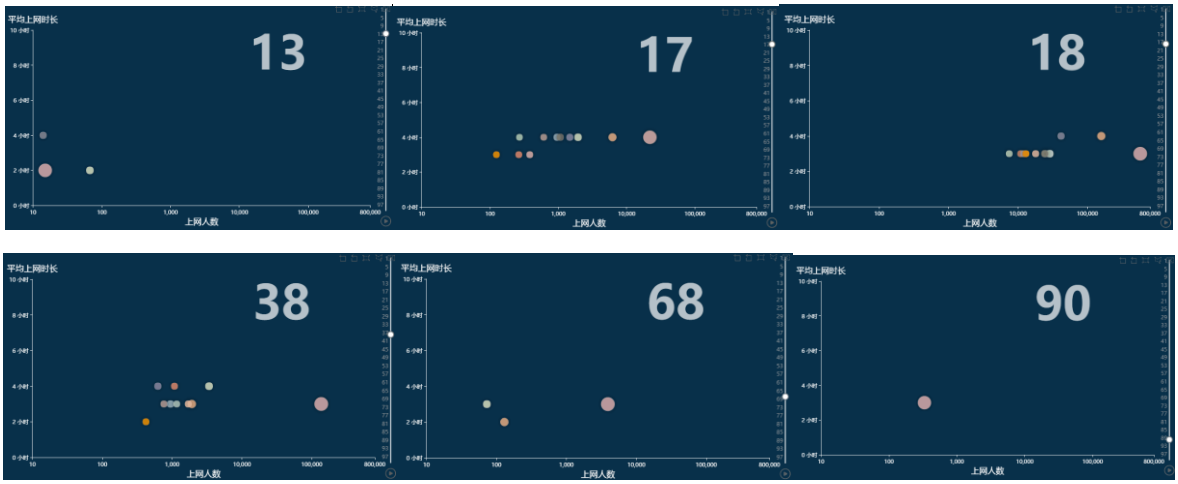


图 4.3 上网记录日历图

其中，右上角的数字代表年龄，不同颜色的圆点代表不同外来人口的省份（如图 2.2 所示），横坐标代表上网人数（取对数），纵坐标代表平均上网时长。我们可以发现虽然有未成年人上网的记录，但是人数不多，当年龄达到 18 岁后，上网次数大量向右偏移（由于是对数坐标，想右偏移是以 10 倍的速度增长），上网人数激增。当年龄超过 30 岁之后，上网人数慢慢回落，60 岁之后，人数急剧减少，80 岁之后，伤亡人数接近无。对于 60 岁以上用户，可以做特别调查，因为通过常理判断，这些很可能是假身份证或“被套牌”的身份信息。

挑战 2.5：根据你所搜集的信息以及分析的结果，试着对某市公安局提出综合性建议。（建议参赛者回答此题文字不多于 500 字，图片不多于 5 张）

综合搜集的信息和分析的结果，我们提议公安局加强对学校周边和商区周边网吧的监督力度对直接接纳未成年人的网吧进行严厉处罚，并对非法利用成年人信息帮助未成年人上网的网吧（如图 1.3 所示）进行调查，并作进一步处理。

从我们对非法经营网吧的分析中来看，北城天街周围直接接纳未成年人的网吧最为密集，如下图 5.1 所示。

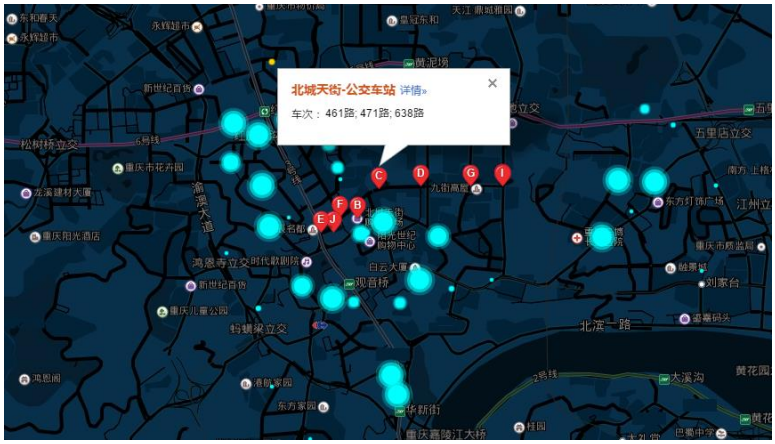


图 5.1 北城天街周边违法经营网吧

从搜集的信息中，我们也看到了如图 5.2 这样的报道“位于江北观音桥北城天街旁的**星网吧，从去年初开业以来，生意一直不差。而这所谓的“生意”背后，却暗藏不少玄机：每到周末，网吧就坐有十来岁、背着书包、穿着校服的小学生。”

图 5.2 未成年学生沉迷在网吧游戏中

根据我们对于上网高峰时间的分析，建议公安局在每天下午 4 时放学/下班之后，加强对商业区和学校周围网吧的查处力度，并引进先进的网吧刷卡系统，如未成年人身份证不能刷卡上网；成年人连续上网超过 5 小时候需要再次刷卡重新上线；对于 60 岁以上的老年人用户限制上网时间等；在防止套牌身份证的同时，也能保证用户们的身心健康。

减少非常经营的网吧数量除了公安部门和网吧营业者的共同努力，还需要全社会的帮助。增强网络安全教育，建立举报机制，对青少年宣传正确的思想价值观念也是减少违规上网行为必不可少的部分举措。