重庆市网吧可视分析系统项目报告

吴嘉骜 21307130203

主要工具：Python 3.12.1、D3 v7.min.js、Plot.js、百度地图 API

一、数据清洗与预处理

本题目数据量十分庞大，存在一些无效数据、脏数据，需要对这一部分进行清洗。同时，为了后续可视化编程方便，需要对数据进行预处理。这部分代码详见 preprocess.py。下面是主要的步骤：

首先，导入数据，查看并转换数据格式。我们使用 Python 的 Pandas 库进行数据处理，查看网吧信息和上网记录表格的字段数据类型，并将相似的列做类型上的统一。主要包括：（1）将 ONLINETIME 和 OFFLINETIME 转换为 datetime 类型；（2）BIRTHDAY 转换为 date 类型；（3）将 AREAID 转换为 int64，和 SITEID 一致。

其次，清洗非法数据。（1）清除有空值的行，如经纬度存在一些“#N/A”符号的记录；（2）删除 CUSTOMERNAME 和 AREAID 列中第一个字符是“`”的记录，否则这对于后续匹配地理位置会产生较大干扰；（3）清除格式不合法的字段对应记录，如非法地理位置、籍贯代码异常等。

最后是预处理数据，主要有下面几项：

（1）删除上网记录在网吧表中没有对应 SITEID 的记录，同时也删除没有上网记录的网吧 ID。这主要是为了可视化编程时减少两张表对应错误或冗余的情况。

（2）从生日中提取出流动人员在上网时的年龄 AGE（2016-出生年份），同时清除计算出的年龄不在 0-100 岁之间的数据。

（3）OFFLINETIME- ONLINETIME，计算出上网时长（以小时计），同时清除上网时长为负数的数据。

（4）查询 AREAID 对应的籍贯名称（省、市、地区的名字），整合到数据中，为流动人口分析做准备：

- 提取 AREAID 的前两位，在 provinces.csv 中查找对应的省份名称，记录到 PROVINCE 列；新建一列“ISLOCAL”，将前两位为“50”的数据标记为“1”，即本地户口。
- 提取 AREAID 的前四位，在 cities.csv 中查找对应的城市名称，记录到 CITY 列。将前四位为“5102”的数据的 ISLOCAL 列标记为“1”。
- 对于整个 AREAID，在 areas.csv 中查找对应的地区名称，记录到 AREA 列。
- 剩下记录的 ISLOCAL 列数值标记为“0”。

二、接纳未成年人上网情况

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。

首先，判定未成年人上网记录，同时判定非法经营网吧。这部分代码见 prob1.py。主要的步骤如下：

（1）AGE<18，即被判定为未成年人直接上网，标记 UNDERAGE=1；

（2）单次连续上网时间过长，不合常理。由于网吧合法经营时长连续不超过 16h，我们认定上网记录中时长>24h 有嫌疑，标记 UNDERAGE= 2；

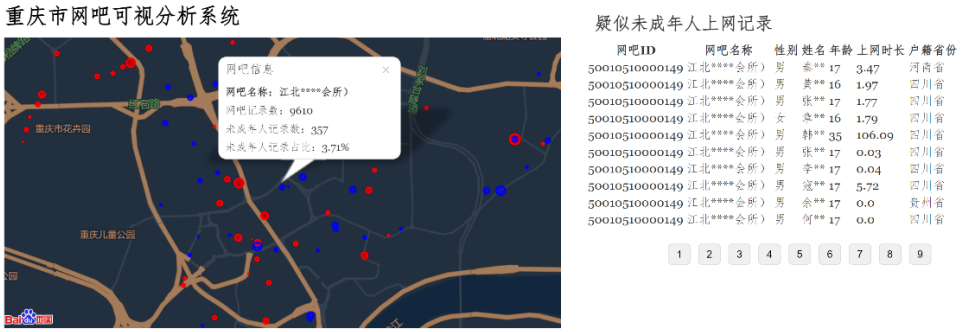
（3）就年龄因素而言，超过 50 岁且单次上网>12h，也认为存在异常，标记 UNDERAGE= 2。

（4） 统计各网吧直接接纳未成年人次数和疑似次数，前者>0 或后者>20，判定该网吧为违规。

接下来进行可视化展现，见 prob1.html：

（1）我们用百度地图 API 展示了重庆市地图，并在图上标记了各网吧的位置，用蓝色表示直接接纳未成年人的网吧，也即“黑网吧”，用红色表示正常营业或嫌疑不大的网吧。点的大小由该网吧的上网记录数决定，记录数越多，点半径越大。如下图所示：

（2）鼠标略过某网吧点，可以展示该网吧的名称、上网记录数和接纳未成年人的记录占比，点击该点可以在地图右侧看到此网吧的（疑似）未成年人上网记录数，可以翻页查看，如下图所示：



通过数据分析，我们发现大约有 20%左右的网吧直接接纳了未成年人，但是接纳的记录数只占了总次数的约 0.5%。

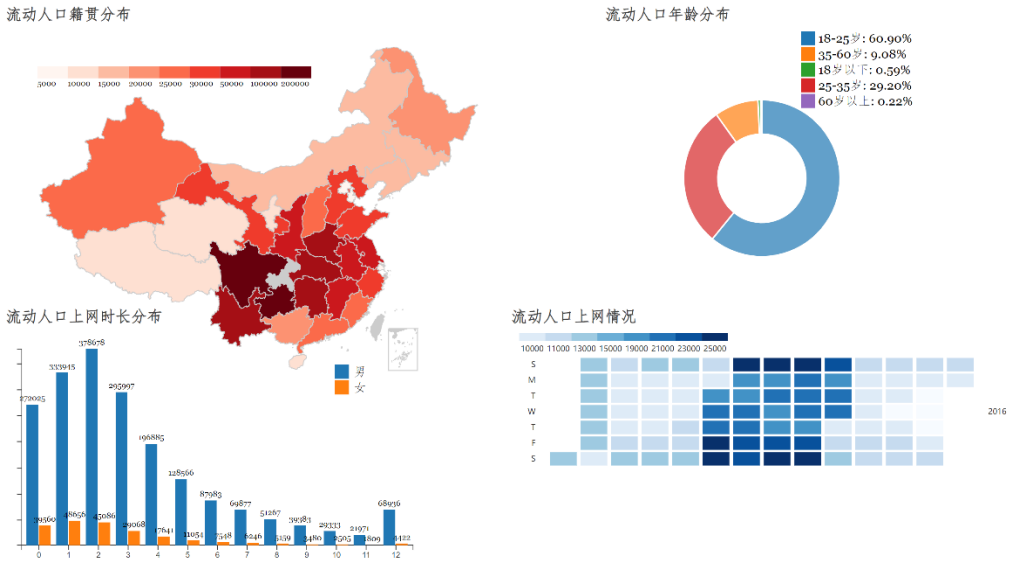
三、流动人口行为分析

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（可从上网时间、时长、上网人员籍贯等维度分析）。

在前面的数据预处理中，我们已经标记了流动人口（ISLOCAL=0），此处说明一下原因：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。所以我们标记 AREAID 是 50 开头和 5102 开头的记录为流动人口。

为了后面展示流动人口上网情况，我们提前获得了每天的总上网人数、每日每小时的上网记录数，见 prob2.py。

可视系统参见 prob2.html，初始展现全部数据的统计结果：



我们可以观察到如下特征：

（1） 通过流动人口籍贯分布热力图，看到流动人员的数量与该省份离重庆的距离有关系（鼠标略过可显示来自该省份的流动人口数），四川省最多，其次是贵州省、湖北省。

（2） 通过流动人口年龄分布环图，可以看到流动上网人员的年龄主要集中在 18-35 岁青壮年，极少数 60 岁以上的人，可能是被冒用身份证上网。

（3） 通过流动人口上网时长分布柱状图，可以看到不论男女，上网时长多为 1-5 小时，且男性上网总时长明显高于女性。

（4） 通过流动人口上网情况日历图，可以看到流动人口的上网高峰日期是每周五、周六和周日，集中在十一假期、十一月份。

点击地图上某省份，可以更新后几个视图，展示来自该省份的流动人口上网特征，和总体相差不大。

上面的特征也很容易得到解释。流动人口中青壮年可能因为缺乏归属感，常去上网，而休息日、假期是最频繁的时候，年末事务繁忙则减少频次。

四、社团发现

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。

我们青年社团挖掘的步骤如下，见 prob3.py 和 prob3.html。

（1）筛选出上网活跃青年—— 年龄在 14-35 之间、且上网记录数>6 的人，认为其有嫌疑，得到待挖掘社团的人员名单；

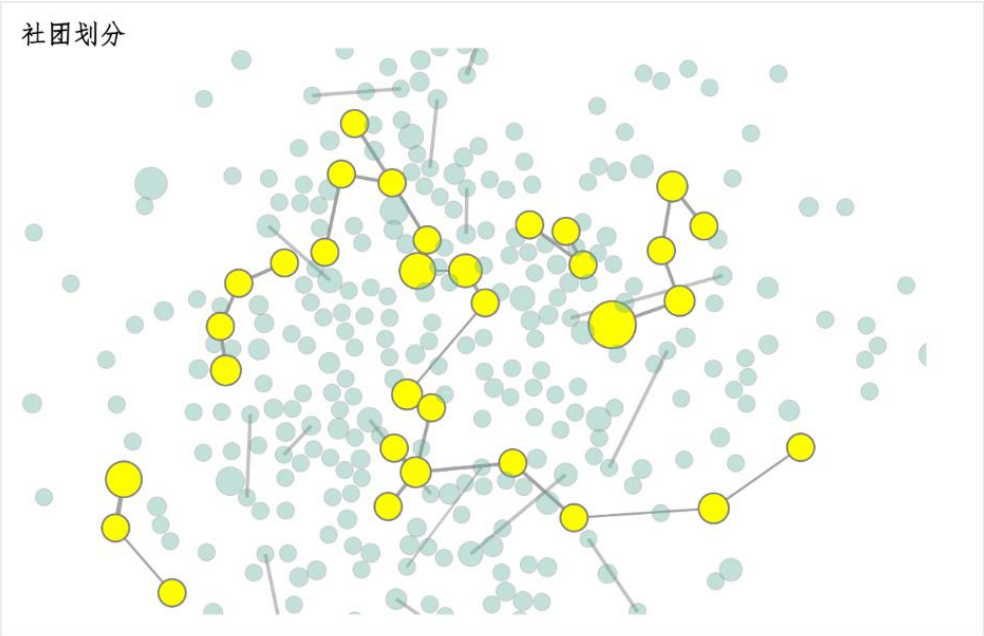
（2）按照上面得到的名单筛选原上网记录，将得到的数据按照上线时间升序排列；

（3）计算相似度：设定滑动窗口大小为 100，对窗口的第一个记录，计算其与后面 99 条记录的相似度，相似度计算方法为：比较上线时间和下线时间差，有一个在 15 分钟以内赋予权重 0.5，若均在 15 分钟以内，则权重为 1，进一步对比是否在同一家网吧，若相同权重再加 1，否则权重不变。

（4）创建图，节点为嫌疑人员 ID，边表示可能为社团关系，当且仅当两人相似度大于 0 时连边，权重设置为相似度。

（5）可视化展现：绘制力导向图，为分析方便仅绘制了权重>0.5 的边（仅有一处相似并不能说明是社团），找出所有连通子图并标记为黄色，点的大小由上网记录数决定，边的粗细由权重决定。

划分结果如下所示：



可以观察到 7 个社团，鼠标略过点可以看到 ID 和上网记录数信息，我们将挖掘得到的社团名单放在了附录中。

五、用户画像

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。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。

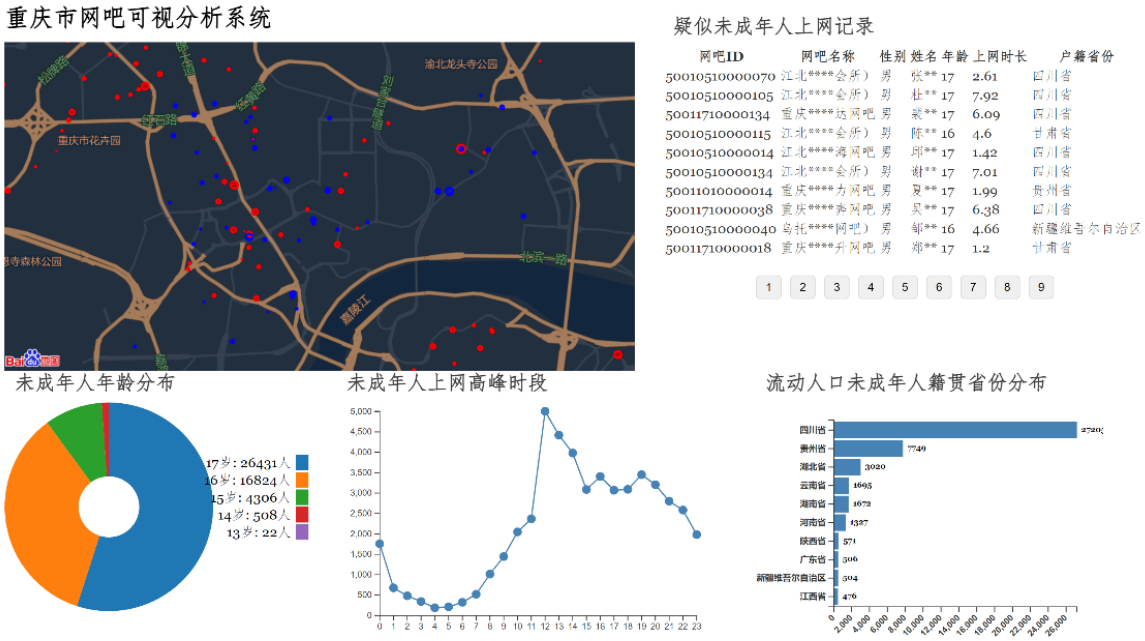
我们主要对未成年人做用户画像。这里直接筛选了身份证上年龄<18 的记录，由于记录数量已经十分庞大，足以使我们得出一些有趣的结论，所以就不再考虑疑似冒用成年人身份证上网的未成年人。

下图是默认展示的完整数据的分析结果，展示了未成年人诸多特征，见 prob1.html。我们归纳总结如下：

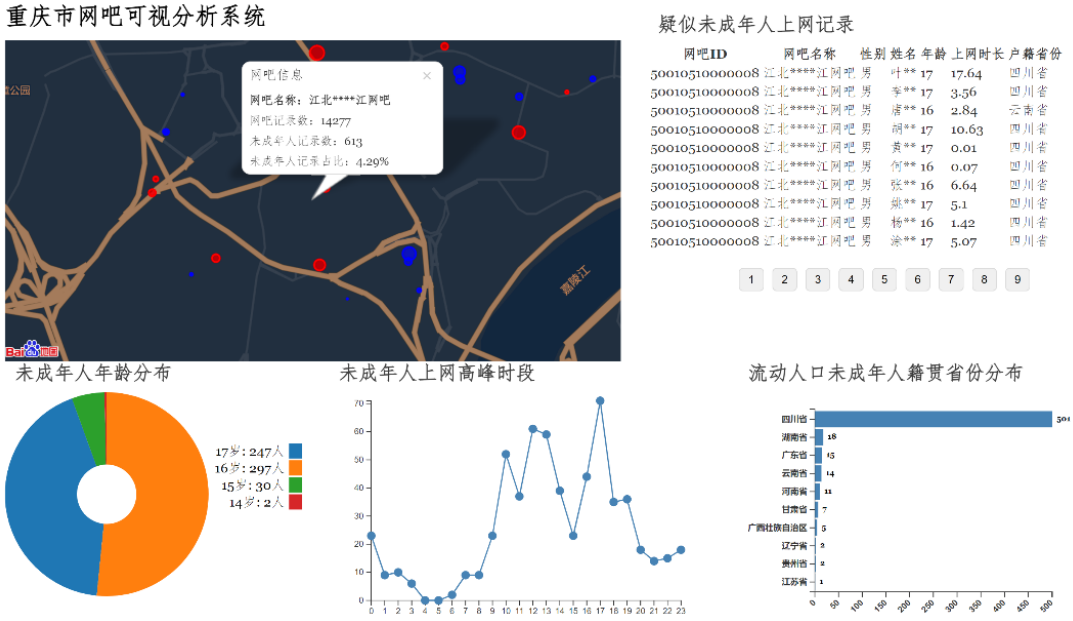
（1） 年龄分布方面，以 17、16 岁居多，即接近法定成年年龄。更小年龄的上网记录也个别存在；

（2） 未成年人上网高峰时段在中午 12:00-14:00，18:00-23:00；

（3） 未成年人流动人口籍贯仍然呈现由近到远分布的情况，四川省占绝大多数，其次是贵州省、湖北省。



点击地图上某网吧点，还可以展示该网吧记录的未成年人画像，举例如下：



我们可以看到，该网吧中未成年人年龄分布和总体很类似，上网高峰集中在傍晚 17:00-19:00，流动人口中绝大多数来自四川省。其他网吧的情况也差不太多。

此外，我们直接放大地图观察，可以发现常接纳未成年人上网的黑网吧的一些特点：周围有学校、商业中心等。

这些用户画像可以得到一些解释：许多户籍来自周边省市的未成年人，在年龄接近 18 岁时，已经有比较强的自主意识，网吧也放松了对他们的约束，接纳他们来上网。这样的网吧多是开在学校和繁荣商场周边，客流量较大，较能得到未成年人的关注。中午、下午下课放学后，是他们上网的高峰，有些甚至一直打游戏到半夜凌晨。

根据上面几个问题的分析，我们可以向公安局提出如下的一些综合性建议：

（1） 建议在中午 12 点、下午 16 点下课放学后，加大网吧的查处力度。同时特别关注学校、商场周边的网吧；

（2） 引进更加先进的网吧刷卡系统，要求未成年人身份证不能进入；

（3） 对于成年人连续上网，达到一定时间（如 6 小时）需要重新刷身份证续时；

（4） 限制老年人口连续上网时间，保护其身心健康，减少冒用身份证风险；

（5） 对长期上网的流动人口（社团）保持一定关注，动用社会各力量对青少年进行正确思想的宣传教育，同时督促网吧合法合规经营，不通宵营业，拒绝接纳未成年人，一经发现严肃查处。

附：第三问挖掘的社团结果

PERSONID	Groupid
b9949157d649b5597d	0
4bfd4bf79f020e8037	0
d5e8475d79ba22db4d	0
9704cd8d8396d600b9	0
f930c19f70e61d5859	0
6b8221a5f344921c29	1
4842d1ae51f6c46d8f	1
6757b375189673ccdd	1
379a8256445d6003b3	1
68f1d030f0f26aa74c	2
a0b2afd876db9f5156	2
3a9c859b567a27bffc	2
f8d9cfc961000151ed	3
ce181ddc14ff619492	3
b452c2cfa1e0674d05	3
d7be9da83fc4dde664	3
9e37870d2602ce1d5e	3
12cf764a2d8748a683	3
0c0881b307d6f3a696	3
d31fece4948c53f66c	3
6ec65de694f586bd57	4
d3317f7f17babfb3a3	4
274599fc60867df589	4
8810a757d4c1c75fbc	4
0575fd4612462a1eb1	5
fcd3f0fda93b391539	5
d467960ceda8f87018	5
a60742cc6d90a4cf5f	5
5acd951389f90d3b96	5
ae8d886215c7bcf9f4	6
7445db6db26c022727	6
a23ba4cd3308ed932b	6