**2017年第三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）**

**背景介绍**

随着网络宽带进入千家万户，网吧这一曾经炙手可热的经营项目，现已举步维艰。于是，一些网吧经营者不惜采用接纳未成年人、通宵营业等违规经营方式来吸引顾客。某市正在开展严查有违规经营的黑网吧的专项执法行动，但网吧数量庞大，传统执法过程，即派遣专门人员挨个检查网吧，不仅耗时耗力，还为黑网吧提供了规避查处的准备时间。为了提升全局掌控能力和执法效率，某市公安局决定采用先进的数据分析技术，从网吧上网登记数据中识别不同上网人群，探索上网人群的时空行为特征，检测非法上网行为与团伙上网行为，为寻找黑网吧提供线索，为犯罪预防和维护社会公共安全提供新思路。作为数据分析爱好者的你，请根据某市公安局提供的网吧上网记录，通过可视分析及其它数据分析方法，帮助分析黑网吧生存现状并解决管理和执法中遇到的现实问题。

**数据说明**

本次比赛提供市内3000多个网吧基本信息及2016年三个月约1600万行上网记录，数据大小为1.7G（压缩前），具体包含网吧基本信息表和上网记录表。网吧基本信息是网吧开业时在公安机关备案提交的，包括网吧编号、网吧名称、网吧的位置信息经纬度等。网吧开业后，会陆续有人来上网，任何一个人上网都需要出示有效证件，网吧登记人员会实名登记上网人的基本信息，并记录他们上线和下线时间。数据中可能存在一些脏数据（例如经纬度为空或无效）需要参赛者注意并处理。两个数据表具体字段说明如下：

网吧基础信息表：

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 类型 | 含义 | 解释 |
| Siteid | String | 网吧编号 | 主键 |
| Title | String | 网吧名称 | 名称已脱敏处理 |
| Lng | Double | 经度 | 根据实际位置进行了随机偏离 |
| Lat | Double | 纬度 | 根据实际位置进行了随机偏离 |

上网记录信息：

|  |  |  |  |
| --- | --- | --- | --- |
| **列名** | **类型** | **含义** | **解释** |
| PersonID | String | 上网人编号 | 根据个人证件号脱敏后得此编号 |
| Siteid | String | 上网场所编号 | 引用网吧基础信息表中的siteid |
| CustomerName | String | 上网人姓名 | 姓名已经做脱敏处理 |
| Xb | String | 上网人性别 |  |
| AreaID | String | 上网人籍贯编号 | 与身份证前六位地区编码一致 |
| Birthday | date | 上网人出生年月日 |  |
| Onlinetime | datetime | 上线时间 | 精确到秒 |
| Offlinetime | datetime | 下线时间 | 精确到秒 |

**题目说明**

**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。

**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）。

**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。

**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。

**问题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。