QG 工作室暑期实习生两日结

姓名： 许继元　　组别： 数据挖掘　年级： 19 级　　日期： 2020.08.13

生活小记	8.12 日，上午的计划是完成后台组给的需求——广州市行政区内两个采样点的 geohash 码，Python 有 pygeohash 模块来实现，所以任务很简单，找到行政区中心点再加一个附近的点，最后把它们进行 geohash 编码即可。实现该需求后，继续分析出租车流量的统计方法，最后采用降低经纬度的精度来形成以所选取经纬点为期望的高斯分布这一方式来进行流量分析。确定方法后，分析并统计了一周的出租车流量数据，并以 csv 文件的格式发送给后台（静态文件）。下午经过商讨，决定用 json 格式进行传输，这样可以提高效率且是高德地图 API 需要的格式，于是下午修改了一下代码结构，重新生成 json 格式数据发送给后台。晚上后台组又加了各需求，需要获取广州市每个行政区的 5000 个经纬点，于是我通过行政区中心点向外扩散的方法随机生成了 5000 个经纬点，由于扩散规模控制地足够小，所以不会超过所在行政区边界。晚上睡觉前汇总并整理各组的进度文档。 8.13 日，上午前端组负责人发来了出租车流量热力图的测试效果，看起来还不错，不过感觉经纬点有点偏移，后来发现其他组员也有类似情况，首先排除地球板块运动的情况，后面发现 GPS 采用 WGS-84 原始坐标体系。一般用国际标准的 GPS 记录仪记录下来的坐标，都是 GPS 的坐标。但在中国，任何一个地图产品都不允许使用 GPS 坐标，据说是为了保密。因此我们只能转化为 GCJ-02 坐标体系，其又称“火星坐标”。在中国，必须至少使用 GCJ-02 的坐标体系，于是找到了算法来进行坐标体系转换，修改完代码生成新的数据经过测试就没问题了，效果很美观，美中不足的是由于分析流量时经纬点精度降低幅度过大，导致地图上大部分都是大流量，这显然不现实，于是经过调整精度降低幅度以及控制数据量，最后经过测试效果很不错。至此，解决了解决流量数据的经纬点偏移问题。晚上选取了五个算法模型（分别为 MLP、随机森林回归、GBDT、XGBoost 以及 LightGBM）进行训练评估，最后选出最合适的流量预测模型 XGBoost。
学习 开发 比赛 概要	学习上，学习了 pygeohash 模块对经纬点进行 geohash 编码。学会了 WGS-84 坐标体系转化为 GCJ-02 坐标体系的算法。 开发上，解决了解决流量数据的经纬点偏移问题。完成了出租车流量数据的分析统计。选取了五个算法模型进行训练评估，最后选出最合适的流量预测模型。

感想收获	这两天项目的进度还算顺利，一切都在正常运作，需要做的是继续熟悉跟进其他小组的工作，多花时间和精力去想想怎么沟通各小组以及项目的优化。
存在问题（备注）	