## （1）调用高德api/调用hmap接口（hmap联系孙修飞/陈善文），通过订单开始和结束经纬度，获取最短路径，对比订单里程，给出异常订单的推荐阈值

#### 1.1期望结果：

给出差异值的分布，从而确定阈值

#### 1.2 思路：

* 调高德API获得内容、调HMap API获得路径内容
* 使用folium显示图像（核对具体路径，检查错误）
* 最终计算得到实际距离，并与数据库中的里程计距离进行对比

困难解决：

发送json必须是\"（双引号） ，python urlencode打包时只出现\' 需要进行字符串的替换；

注意坐标系之间的转换！

高德：GCJ-02 火星坐标

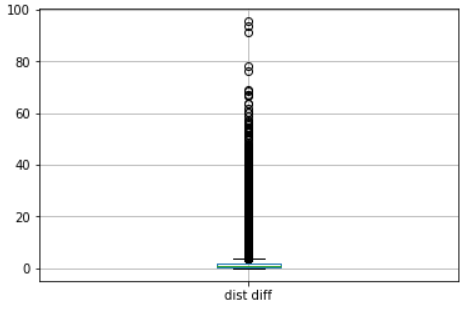
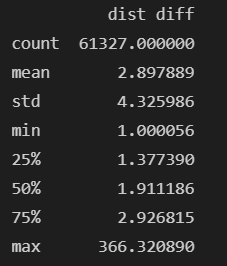
Hmap：WGS84 4326

调用Hmap时，传入的点（高德）故先转为4326，然后得到若干导航路径点，同样是4326的。如果要在高德的街景图上标出，亦要记得转回火星坐标后再用folium画出；

#### 1.3 具体实现：

1. 利用python线程池进行多线程调用Hmap-API，计算实际距离；
2. 考虑到异常路径一般不会出现在起步距离以内的订单中，故仅考虑实际里程大于3km的订单；
3. 筛选出 zke\_lch\_jyqk - Hmap\_dist > 1 （考虑了Hmap与高德Amap本身的导航距离差异）的里程差异分布图（司机比导航多开的路程），画箱线图并统计数值特征；

下图为44万条数据的差异分布图，纵轴为司机比导航多开的公里数。

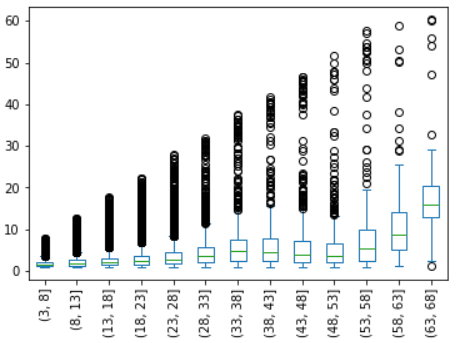
IQR=Q3-Q1，

上限=Q3+1.5IQR = 5.23km

根据异常值上界公式计算出，异常值上界为5.23km。

但是实际上，里程差异与实际行驶里程之间也有一定关联，这样并不能很直观的看出其关联。

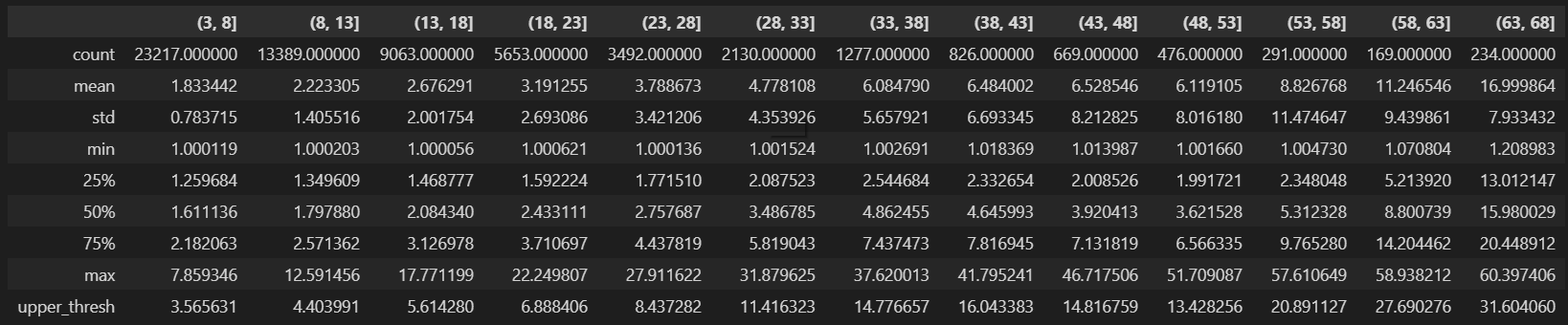
1. 于是，试图将实际里程进行按照5km的间隔进行分段，并画出每一段的分布差异图。



上图为不同行程距离段对应的行驶距离差异分布；可以看到随着行驶距离的增加，其异常上界总体呈上升趋势。所以阈值的筛选也应该采用分段式计算。

下表中，列名为对应的行驶距离段，列里的统计值都是有关司机实际行驶值与导航距离之差的数值特征。

其中，upper\_thersh为计算出的异常值上界（即司机比导航多开的距离），如果司机实际行驶值与导航距离之差大于该上界，则很有可能是异常驾驶订单！



实际数据见：task1\_ods\_july\_only\_with\_hmap\_dist.csv

实际describe信息见：task1\_ods\_july\_diff\_with\_hmap\_describe.csv

## （2）根据订单开始和结束经纬度，通过geohash编码，获取相同起点和终点的订单里程分布，给出异常订单的推荐阈值

#### 2.1期望结果：

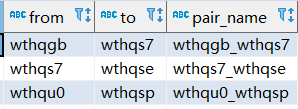
给出不同坐标点的订单里程分布，观察异常值（横坐标订单起始、终点坐标对，纵坐标里程数）

#### 2.2 思路：

1. 先矫正数据，更正经纬度，并计算出对应的geohash值（SQL实现，否则有点慢）
2. 取geohash前6位，就起点与终点进行分类
3. 将相应的geohash对进行合并，如A-B与B-A，其实是同一段行程的订单，只不过出发点和目的地颠倒了，应该需要被合并。

合并方法：

先在python中生成dict字典（将A-B和B-A的value都设置为A\_B！），然后导入数据库中。



然后，为from与to字段建立hash索引;

最后与目标表进行匹配，添加pair\_name字段。

1. 根据需求一的分析，也按照行程距离进行逐段分析，并通过可视化的方式，将坐标对与路线画在地图上，核对是否符合实际情况。
2. 同时通过Hmap计算理想路线，计算方式为先将坐标对通过逆geohash得到区块中心坐标，然后调用Hmap-API进行导航，因为并不是严格与订单的起始、终止坐标一致，故与实际路线有一定误差！

具体数据见：

task2\_all\_pairs\_with\_dist.csv(所有坐标对的订单距离统计，每一列都是一个订单的距离，最后一列”cnt”代表该行空距离的个数，第一行cnt = 0)

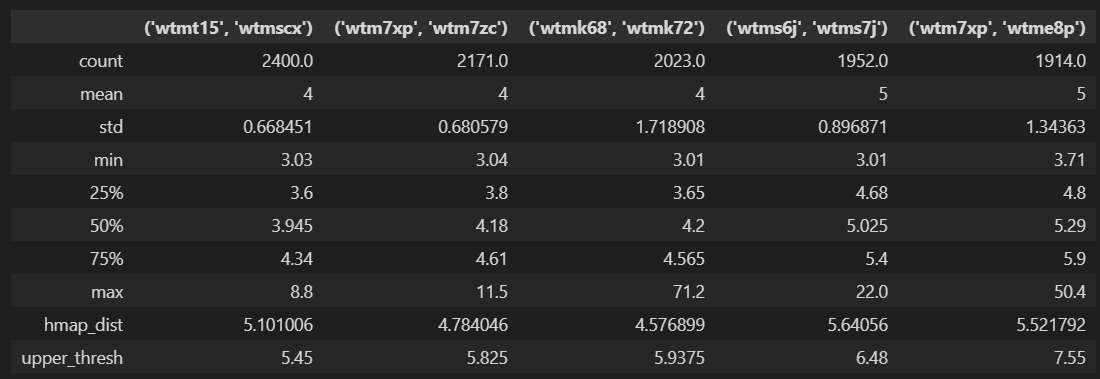
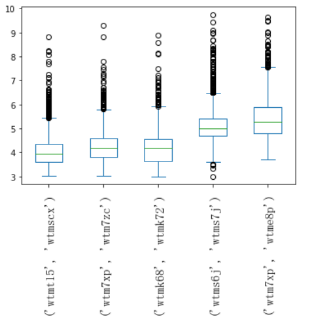


task2\_all\_pairs\_with\_dist\_describe\_with\_thresh\_0903.csv（所有坐标对的describe，包含给出的异常值上界“upper\_thresh”）



#### 2.3 具体实现：

1. 3-6公里的路程（最热门前五条，即订单数最多）



说明：

左图为不同坐标对的行驶里程分布图，其下标为出发地与目的地的6位geohash，纵轴为里程计实际读数；

右图为其数值特征统计表，其中：

hmap\_dist为Hmap导航出来的理想距离；

upper\_thresh为司机实际行驶异常值的上界，即只要实际行驶距离超过该上界，就很有可能是异常驾驶订单！（与需求一不同）

具体路线分析：

（开头的标号为geohash最后两位，起点和终点的geohash区用方框画出，且路线与方框的颜色相同）

该距离的订单一般都是住宅区到周边的工作地点，一般均值距离都在4-5km，所以异常上界可以定在6-7km。

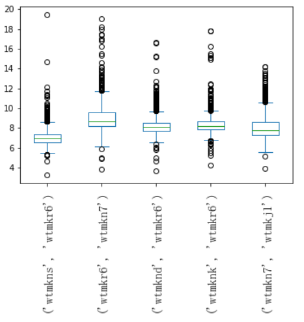
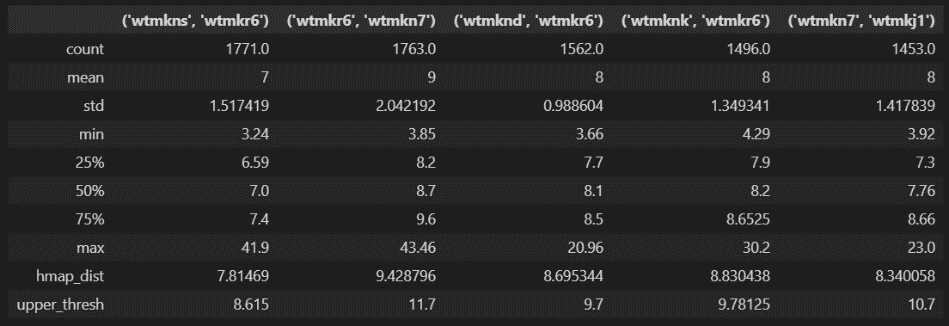
6j-7j : 外包服务大楼---地铁站 68-72： 阿里巴巴---住宅区

xp-8p\ xp-zc: 科技园 -- 住宅区 15-cx：地铁站到住宅区

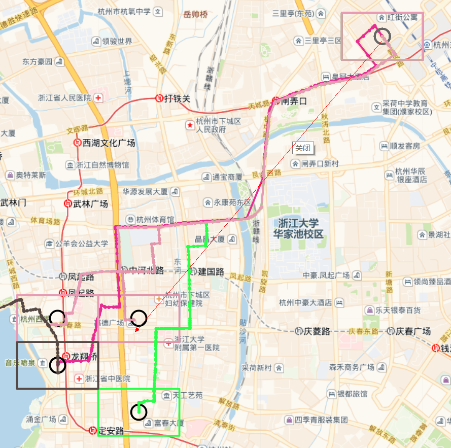
 

1. 6-10KM（最热门5条）

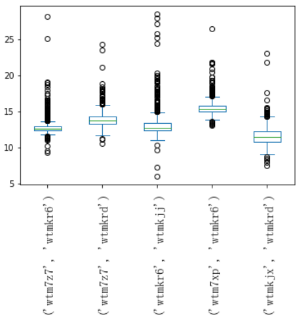
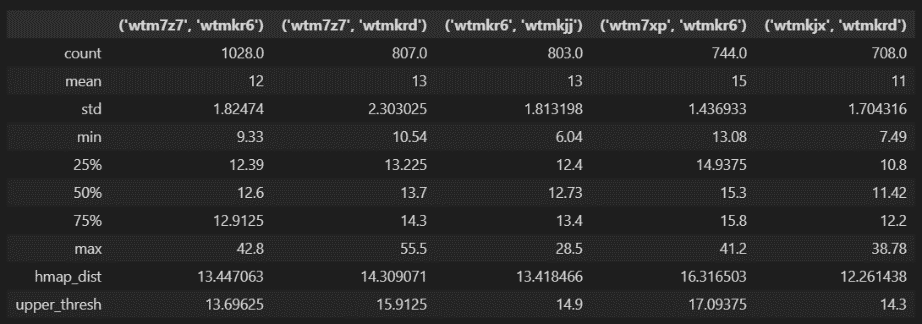
具体路线分析：

n7-j1: 西湖--灵隐寺 剩下四条皆是由杭州东站到西湖周边

该距离的订单一般都是旅游订单，一般均值距离都在8-9km，所以异常上界可以定在10-11km。

1. 10-15km

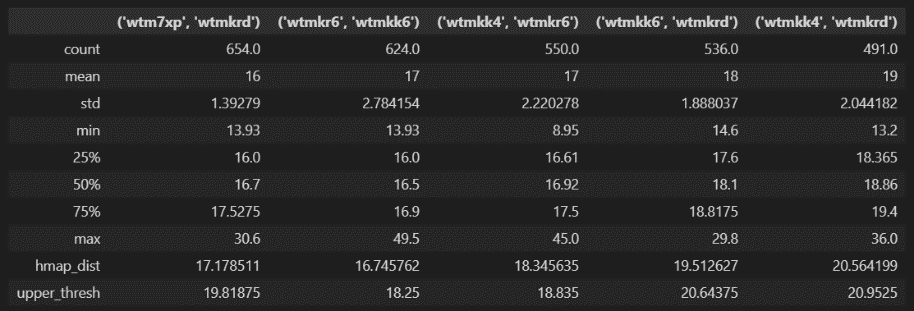
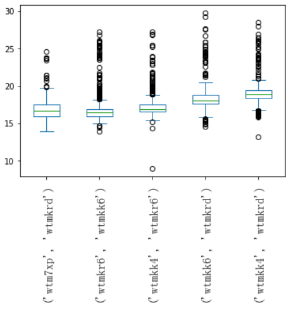
 

Jx-rd 东站到浙大（玉泉、西溪校区）浅色 xp-r6：东站到科技园 灰色

r6-jj 东站到灵隐山下 深色 z7-rd：东站到滨江区人民政府 灰绿

1. 15-20km

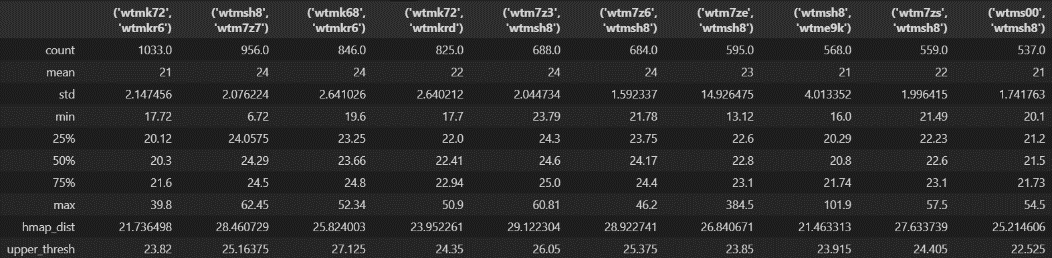
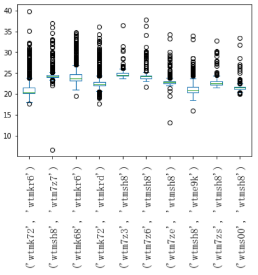


基本上都是东站到紫金港与西溪湿地之间，有一条是到滨江科技园（黄色）

异常值上限可以定在20km；



1. 20-25km （十条最热门）



可以看到，20KM是一个分水岭，低于20KM的订单基本围绕东站，而高于20KM的订单，基本围绕萧山机场。

72-rd 、 68-r6：东站到阿里 9k-h8：杭州南站到杭州萧山机场

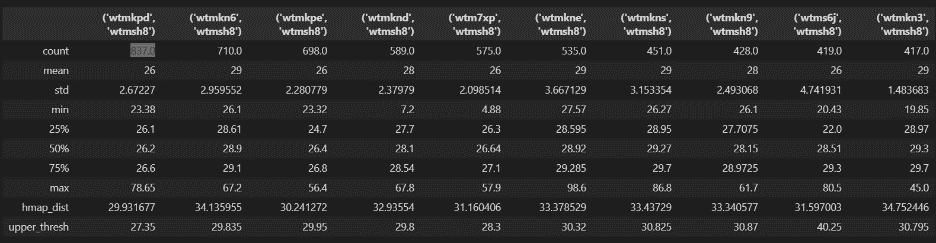
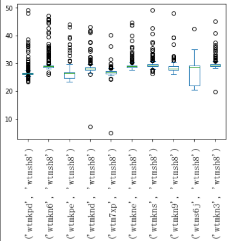
s00-sh8：绿都国金中心（黄色）

其余六条都是，市政府周边到杭州萧山机场



建议阈值可以在25-26之间；

1. 25-30km



6j- h8机场到杭电大学城周边 xp-h8 机场到科技园（黄色）

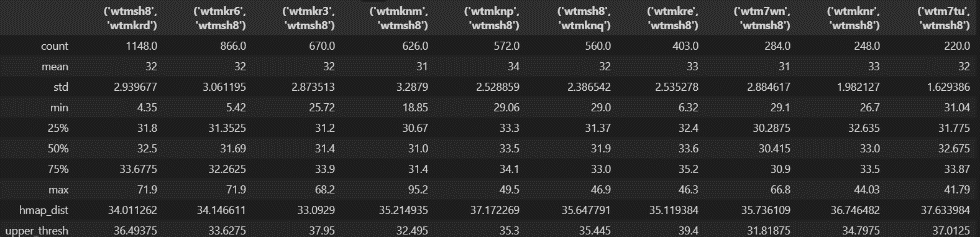
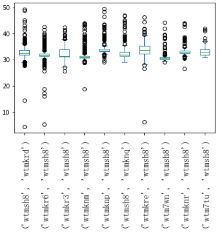
pd-h8、pe-h8 机场到万象城周边（灰色框、紫色框）

其余都是机场到西湖景区周边



建议阈值30-31KM

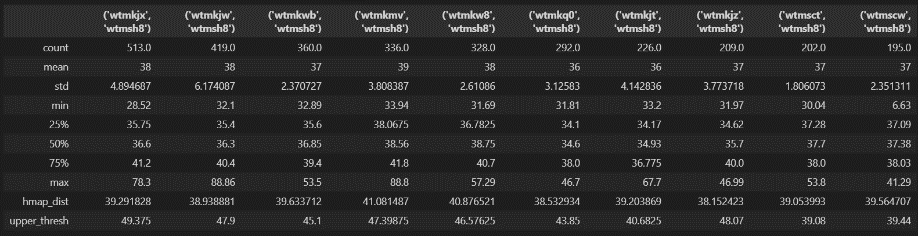
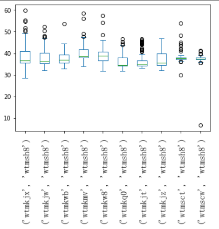
1. 30-35KM

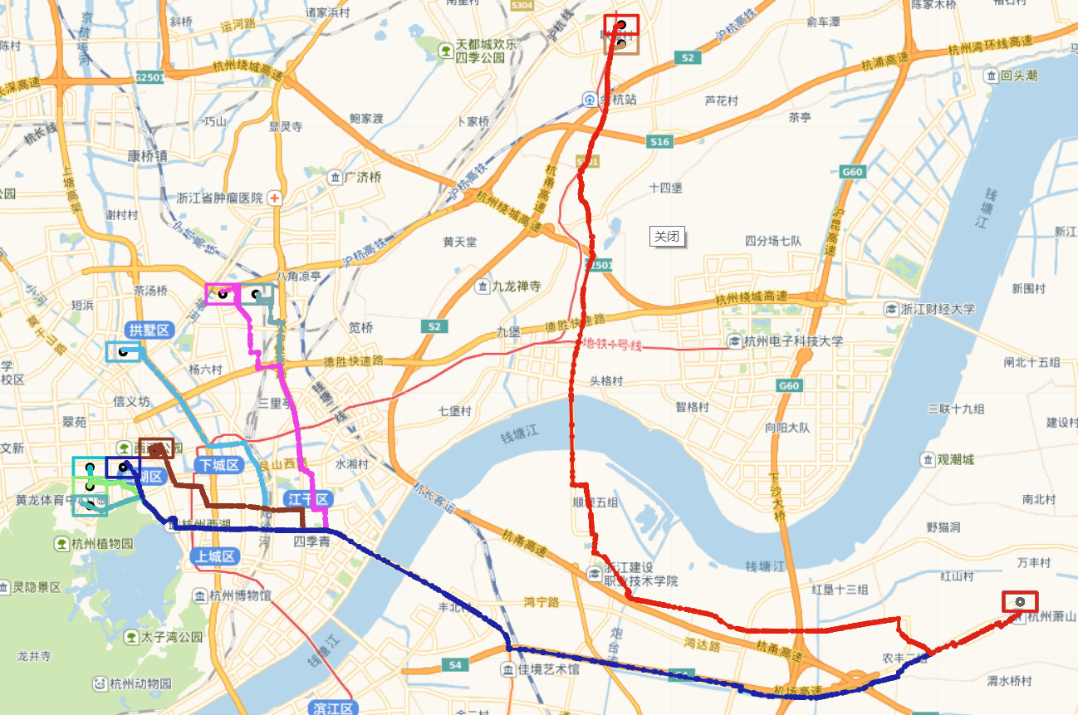




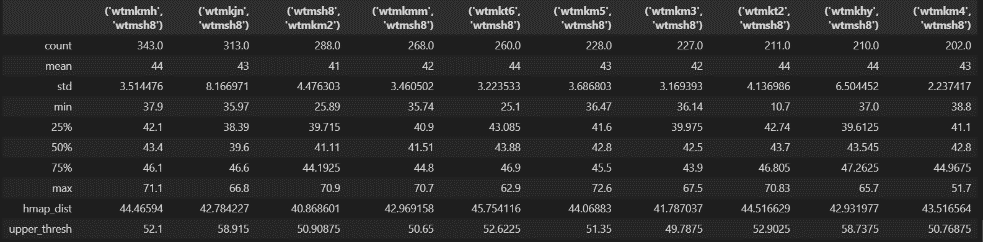
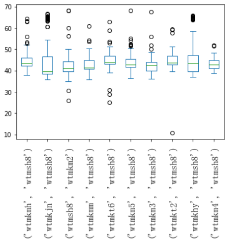
1. 35-40km

基本上都是从萧山机场到浙大玉泉附件的科技园、到余杭区、到拱墅区



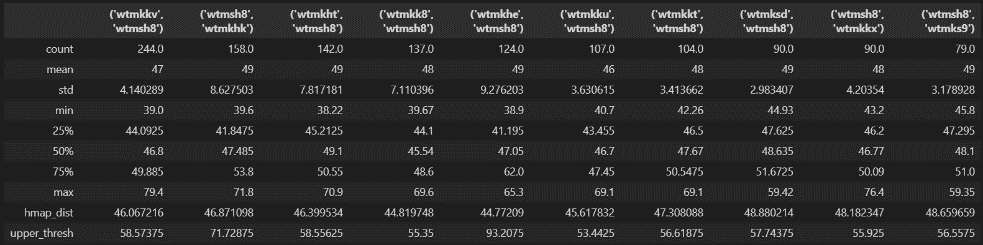
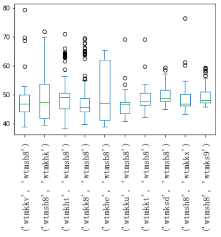


1. 40-45 km



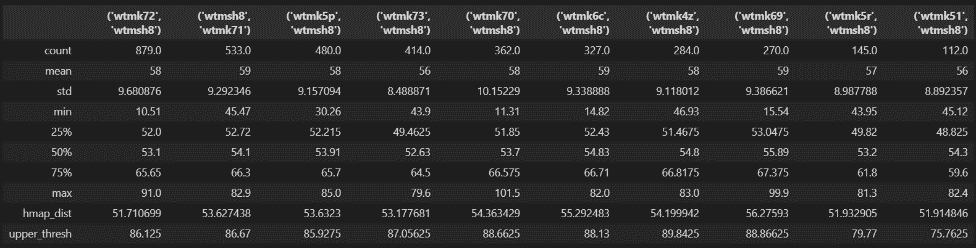


1. 45-50km





1. 55-60km



基本上都是机场到阿里

