【项目09】 中国姓氏排行研究

作业要求

1、数据清洗、整合

要求：

1. 将“data01”、“data02”分别读取，并且合并成一个数据

② 结合“户籍地城市编号”及“中国城市代码对照表”数据，将城市经纬度连接进数据中

③ 分别提取“工作地”中的省、市

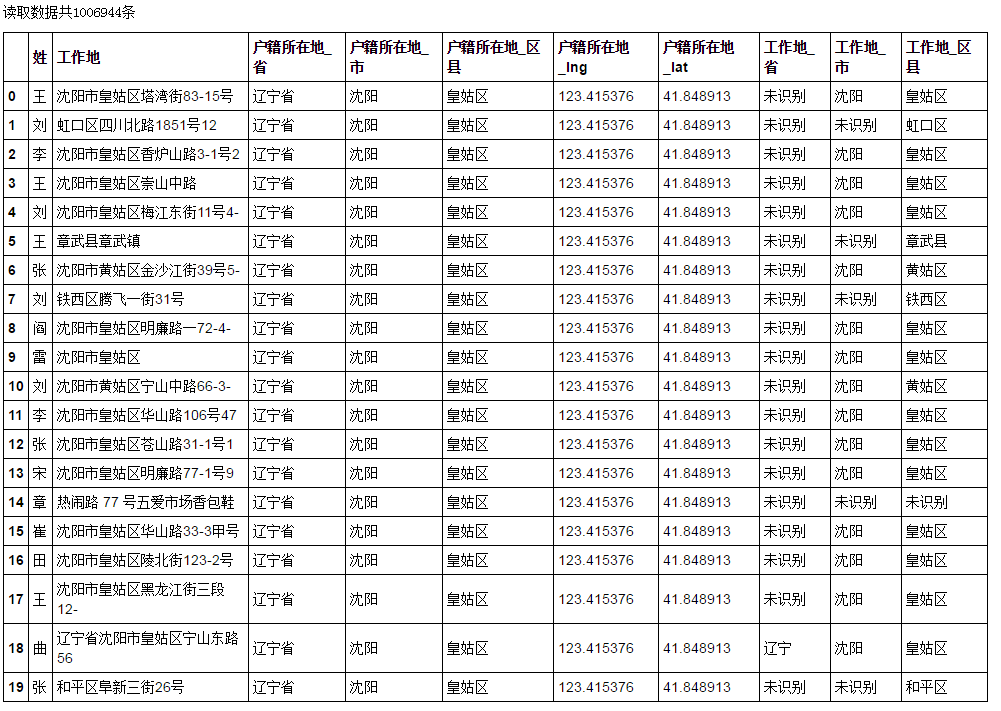
提示：

① 可以先读取“data01”、“data02”，然后用pd.concat()来连接数据

② 新建字段“工作地-省”，“工作地-市”，“工作地-区县”,如果数据中“工作地”字段无法提取省和市，则用“未识别”填充单元格

\* 通过查看识别后的单元格，如果字数超过5则为“未识别”据

② 新建字段“工作地-省”，“工作地-市”，“工作地-区县”,如果数据中“工作地”字段无法提取省和市，则用“未识别”填充单元格

\* 通过查看识别后的单元格，如果字数超过5则为“未识别”

整理后数据大概1006944条

2、查看姓氏“普遍指数”，普遍指数=姓氏人口数量

要求：

① 将数据按照“姓”做统计，找到数量最多的TOP20

② 分别制作图表，查看姓氏TOP20的数量及占比

\* 建议用bokeh出柱状图，并且为联动图表

③ 查看“王”姓的全国分布

\* 这里导出excel高版本文件，用powermap查看，绘制密度图

\* 同时可以尝试用echarts绘制空间柱状图来查看

④ 查看“姬”姓的全国分布

\* 这里导出excel高版本文件，用powermap查看，绘制密度图

\* 同时可以尝试用echarts绘制空间柱状图来查看

提示：

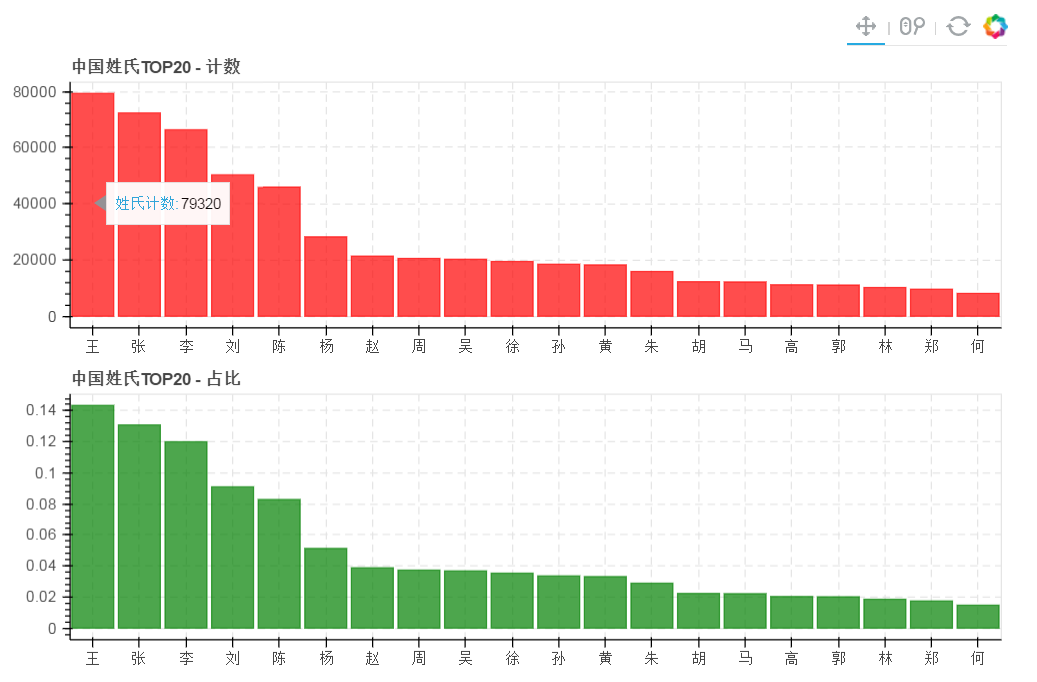
① bokeh中绘制联动图表时用gridplot

② powermap需要office2016的excel才会有，并且必须存储xlsx格式

③ powermap中需要通过在“值”中设置“姓的计数”才能正确显示热力图

④ powermap中可以通过“主题”来调节配色风格 / “平面地图”选项来调整球面可视化或者平面可视化

⑤ echarts绘制图表之前，需要对数据按照“lng”（或者“lat”）汇总，得到同一个地点的该姓氏人数，然后绘图

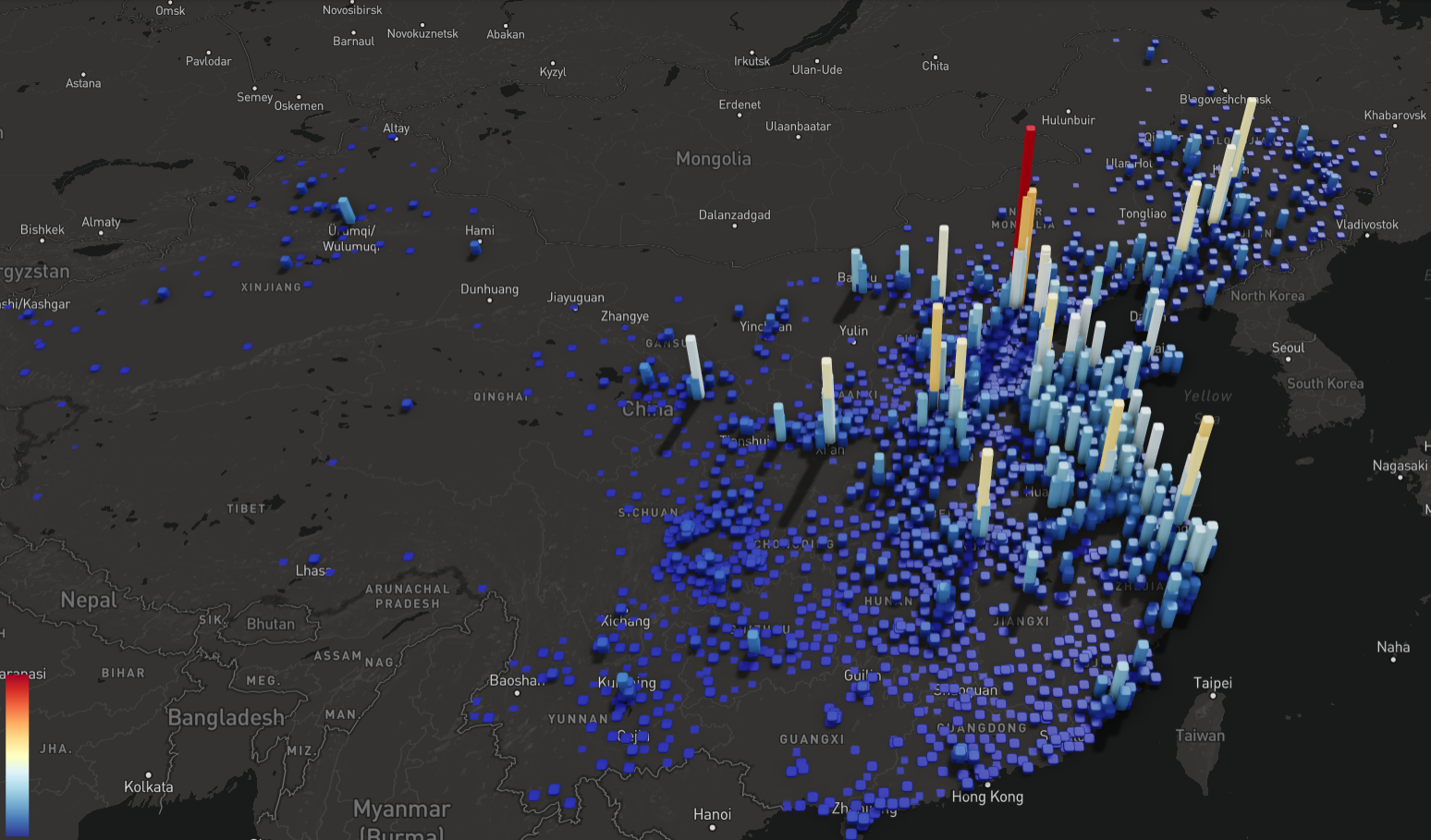
⑥ ecahrts通过设置以下参数来达到效果：视角远近、点柱大小

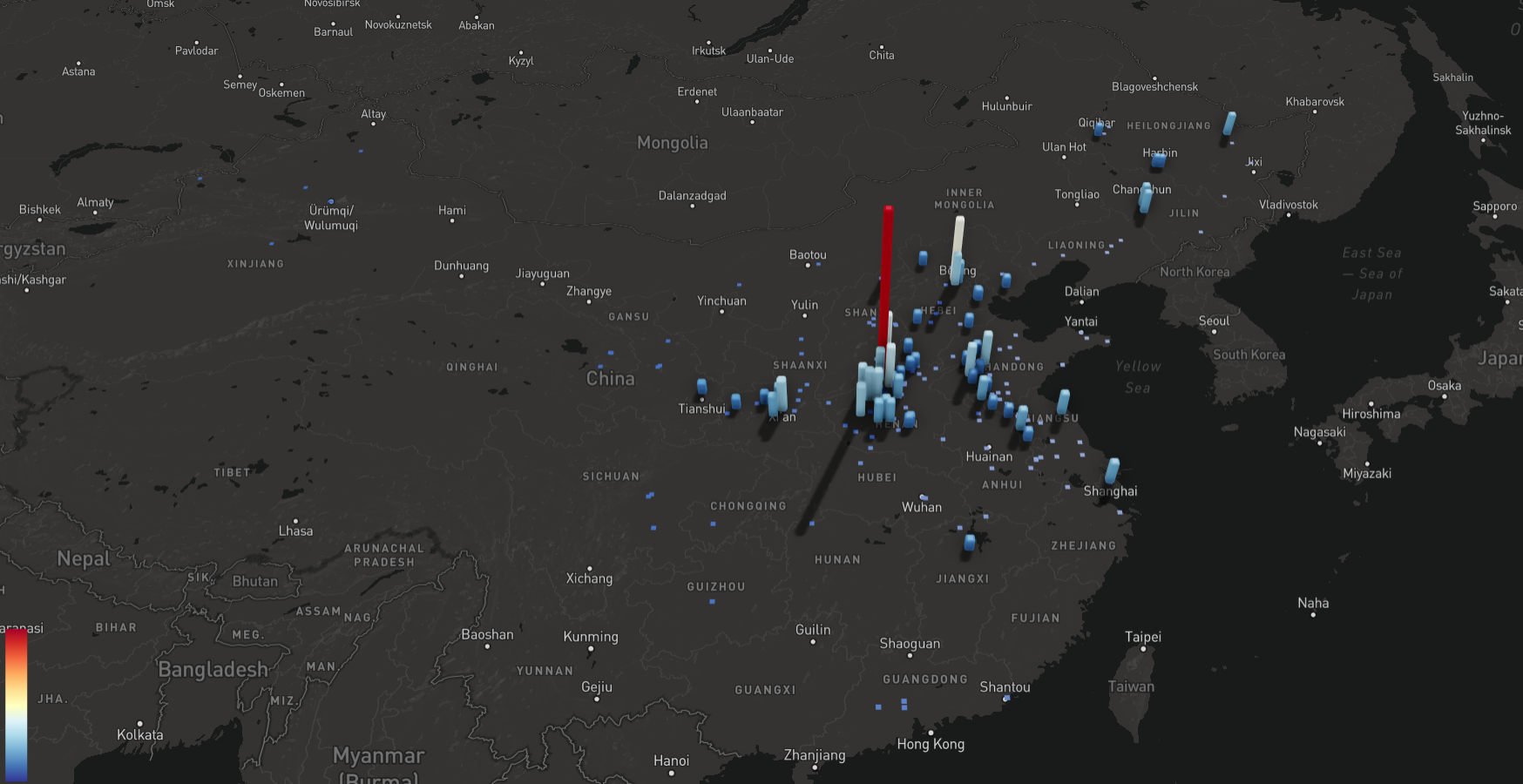
“王”、“姬”姓氏热力分布图：





“王”、“姬”姓氏空间柱状图：





3、查看姓氏“奔波指数”，奔波指数=姓氏人均迁徙距离。迁徙距离为户籍地所在地级市与现居住地所在地级市的距离。

要求：

① 根据识别的工作地，通过Geocoding查询到对应坐标

② 选择一个姓氏，计算并查看其姓氏的奔波指数，并计算该姓氏的人均通勤距离

\* 在python中筛选数据之后，qgis内做空间分析

③ 按照起点和终点做计数，汇总同一个迁徙路径的数据

④ 通过echart制作通勤OD图

\* 可以将生成的line文件导出geojson，用空间线性轨迹图来表示

\* 这里线的valye为该迁徙路径的汇总计数

提示：

① 可以筛选一些好玩的姓氏：汤、朴、廉、何、叶、冉等等

② 需要对数据的工作地进行筛选，其中“工作地\_市”、“工作地\_区县”未识别的数据删除掉

③ 导出数据时，尽量columns名用全英文，避免qgis中出现乱码

④ 计算人均通勤距离的时候，需要删除掉户籍地与工作地相同的人（未迁移的人）

⑤ 在官网metrodata.cn的小工具中找到geocoding

⑥ qgis中需要安装插件“LinePlotter”来转线，并在qgis中计算平均通勤距离（需要投影，投影经度带可选48）

⑦ shapefile转geojson时，注意shapefile文件要投影回wgs84地理坐标系

这里以“汤”姓为例

