拉勾网数据清洗及分析

读取数据

- 加入现在我们获取到了之前爬取好的数据
- 现在需要批量的将数据读入进来
- 思考：一个一个数据读应该如何操作，是否可取呢?

批量读取数据

- 使用os模块获取文件名称列表
- 构造一个循环代码将数据全部读入进来，储存到变量df中

数据清洗

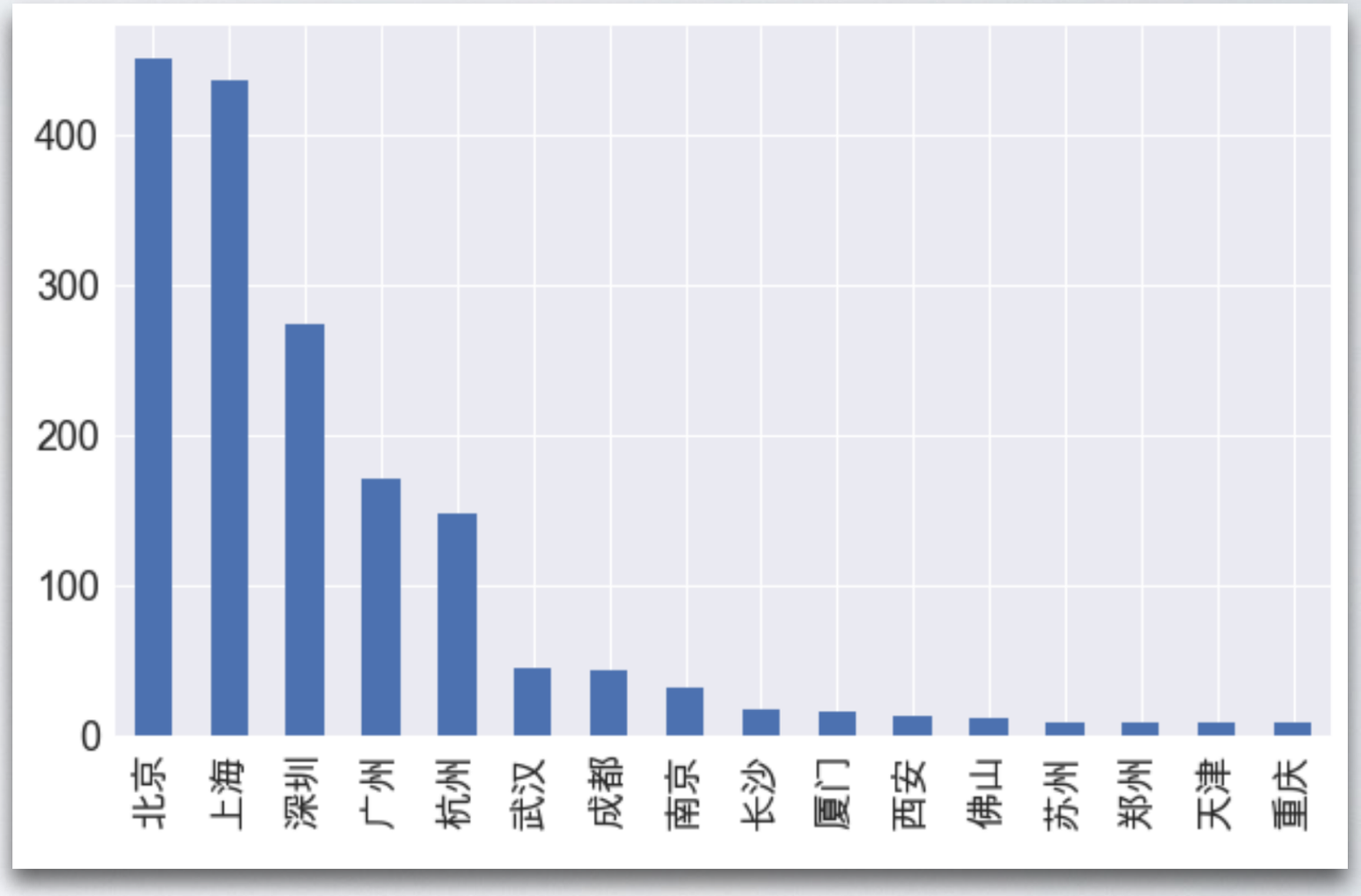
- 首先我们比较关心的是，在全国哪些城市中数据分析岗位的需求量比较高，因此对城市数据做一个统计分析

- 发现没有北京，什么原因呢?

- 将空值数据进行填充，再进行统计

数据可视化

- 将城市岗位数量绘制成柱形图
- 优化绘制图片使其美观



绘制地图分析

- 如果能够更加生动形象的表示出不同地区的不同需求呢?
- 因为城市属于地理数据，所以我们想要将数据绘制成地图，这样可视化效果会更好一些!

PYECHARTS介绍

pyecharts 中文文档

https://pyecharts.org/#/zh-cn/

github 主页

https://github.com/pyecharts/pyecharts

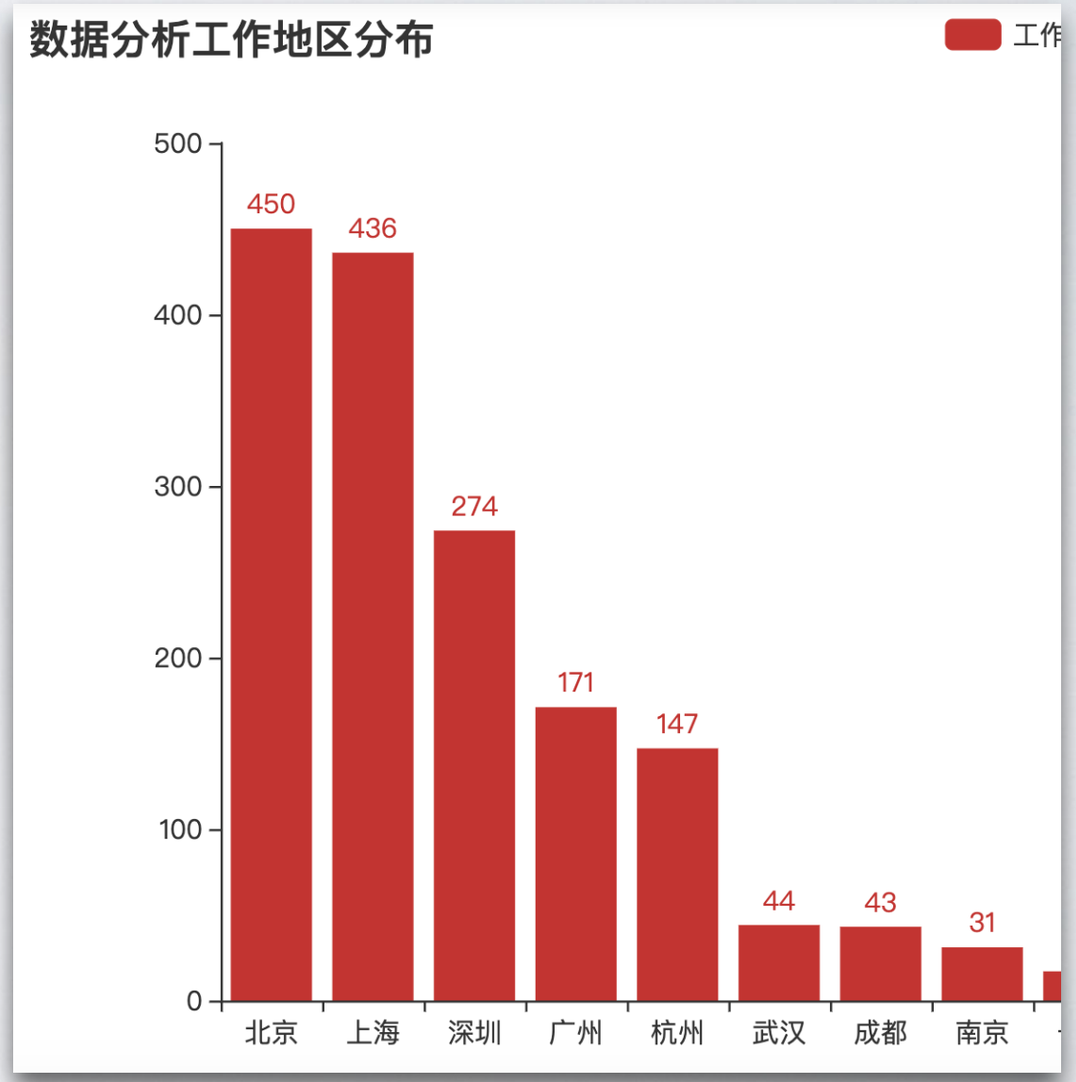
新版本修改说明

https://github.com/pyecharts/pyecharts/issues/1033

使用PYECHARTS绘图

- 我们首先使用pyecharts绘制一个和刚才一样的柱状图

- 熟悉一下这个绘图包的基本操作思路



绘制地理图GEO

首先需要先安装地图包

选择自己需要的安装

$ pip install echarts-countries-pypkg

$ pip install echarts-china-provinces-pypkg

$ pip install echarts-china-cities-pypkg

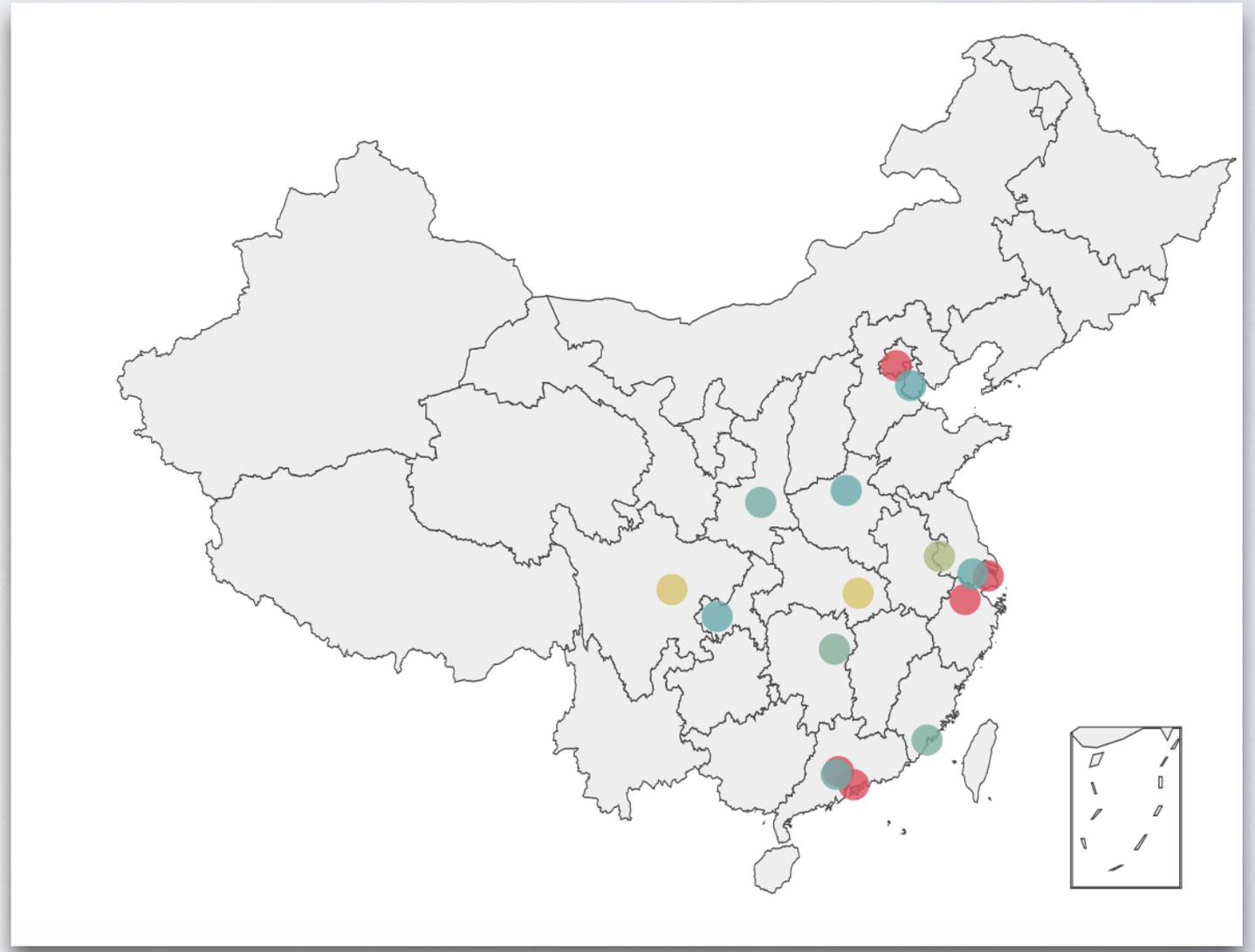
$ pip install echarts-china-counties-pypkg

$ pip install echarts-china-misc-pypkg

$ pip install echarts-united-kingdom-pypkg

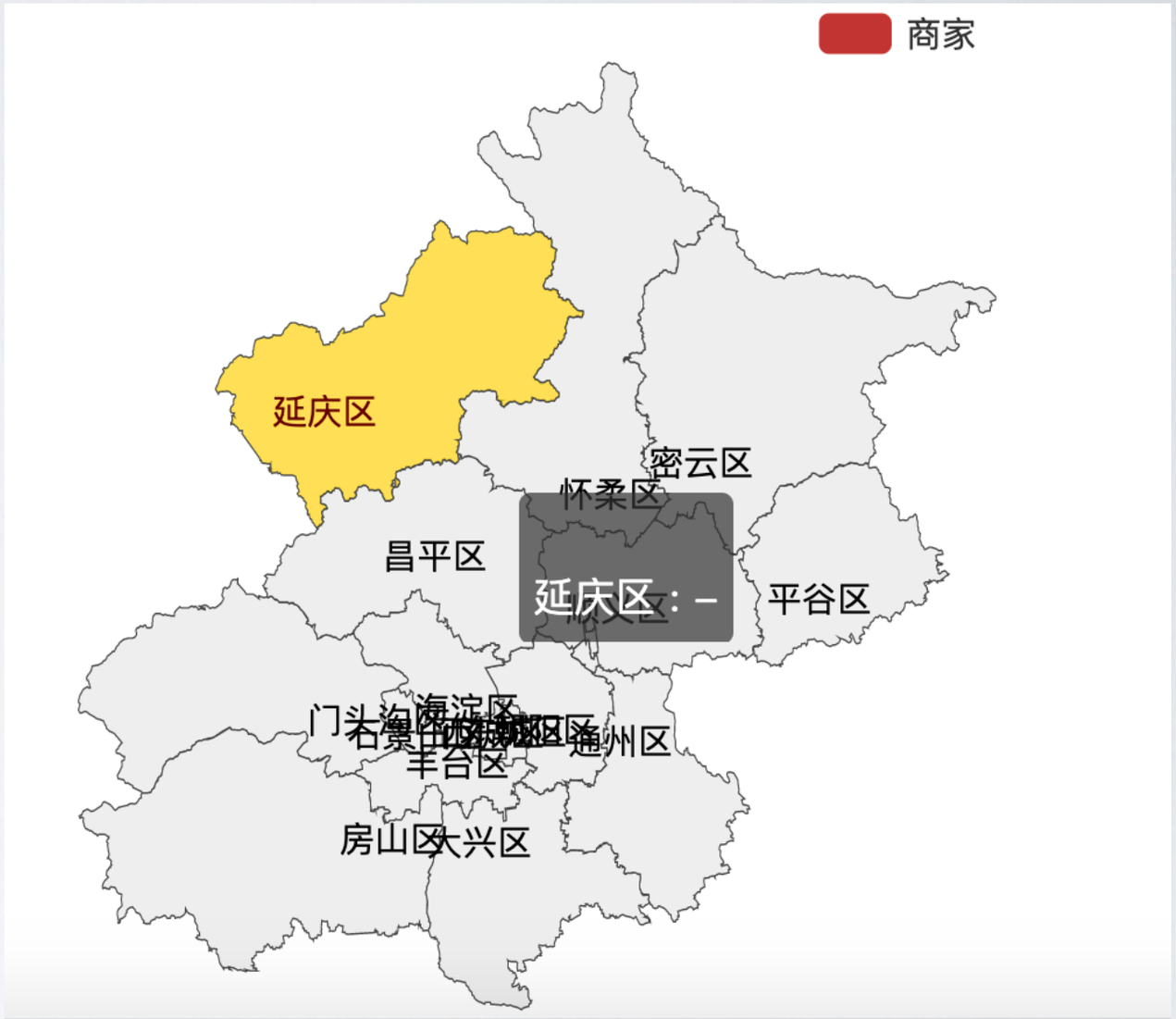
绘制地理图GEO

- 首先把刚才的数据想办法导入进来
- 按照格式编写代码绘制地图
- 这里我们采用连续型标记地理图绘制



地图绘制练习

尝试绘制一个北京的地图
并生成一些数据进行绘制



删除重复值

- 数据中可能有重复值存在, 因为其中有广告信息存在,所以一套招聘信息连续出现了多次
- 检测数据中是否有重复值存在
- 删除数据中的重复值

筛选北京数据

- 单独提取出北京招聘数据
- 把数据中的中括号去除掉
- 尝试能否统计出北京不同区的招聘需求数量

整理工资数据

1. 将工资分裂成两列,最低工资列,和最高工资列
2. 将工资转换成数值型
3. 统计一下最低薪资的平均值
4. 统计一下最高等级薪资的平均值
5. 计算一个总的平均薪资
6. 将各个城市数据进行分组,统计各个城市的平均最高薪资

•