

**《旅游+文本数据分析与挖掘》**

**全国热门旅游景点数据分析+可视化**

|  |  |
| --- | --- |
| 学 院： | 信息工程学院 |
| 专 业： | 计算机科学与技术 |
| 班 级： |  |
| 姓 名： |  |
| 成绩评定： |  |

信息工程学院制

1. **项目背景**

(500字以上)

2021年上半年延续复苏，下半年受疫情扰动。德尔塔变种影响下国内波澜再起，一季度和三季度先后经历了就地过年和限制跨省出游等防疫措施限制，国内旅游人次较2019年同期下降超40%，恢复程度环比低于2020年下半年。整体上2021年休闲服务行业恢复略差于疫情控制较好的2020年。从节假日旅游情况来看，2021年旅游人次与2019年同期水平相比前高后低，旅游收入较2019年同期水平恢复相对较平稳，旅游需求承压但仍较坚挺，长途游比例降低后居民旅游消费能力和水平逐渐提高。

疫情对休闲服务行业带来了不可逆的变化。需求方面，疫情扰动下居民出行习惯逐渐改变，人们逐渐习惯了省内短途出行，在城市周边寻找新的旅游体验。供给方面，疫情倒逼行业转型，随心飞、随心租等产品将有效填充周末及小假期出行需求，进一步推动非节假日短途出游。后疫情时代，旅游行业逐步复苏与海外消费回流将成为休闲服务行业新的增长驱动力。长期来看，我国中产阶级家庭数量提高将逐步带动消费升级，促使旅游需求从满足吃住行游购娱六大基础需求向更深层次的文化旅游、深度旅游过渡，游客将逐步重视旅游体验，倒逼行业中高端化转型。

11月24日，世界旅游城市联合会首席专家、中国旅游协会休闲度假分会会长魏小安发表了题为《2022：胜利在望》的主题文章。他指出，最近这一波疫情反弹已经将近尾声，冬季的候鸟型游客又将聚集在海南云南两广等地，在疫情防控有效的基础上，春节的高潮必然来临，之后就是江南，阳春三月，旅游回暖。

文章指出，最新的专家调查问卷显示，45％的受访专家认为国际旅行将在2024年甚至更晚才能回到2019年的水平。43％的受访专家认为国际旅行将在2023年回到2019年的水平。

11月26日，在中国经济网《“20条”下，疫情防控与旅游发展如何实现双赢》一文中，同程研究院首席研究员程超功坦言，尽管疫情形势依然严峻，但业内普遍对旅游业的复苏抱有信心。“本地休闲度假游、短途周边游等新兴消费需求是支撑行业稳定复苏的重要动力源。”程超功说，随着大众出行的决策周期缩短、出行半径缩小，以周边短途游为主的旅游产品更受欢迎；且局部地区旅游出行政策不确定性加强，大众的出游决策更加谨慎，以小家庭为单位的自驾周边短途游热度持续走高，新的消费需求不断催生产品端的创新、迭代，也带来了更多的市场机会。全国各地的游客也在期待旅游也的“回暖”和复苏。

1. **问题分析**

通过全国各地区景点门票的售卖情况，简单分析一下全国比较热门的景点分布和国民出游情况，看看哪些景点比较受欢迎。解决如下等问题：

1、全国景点分布

2、国民出游分析

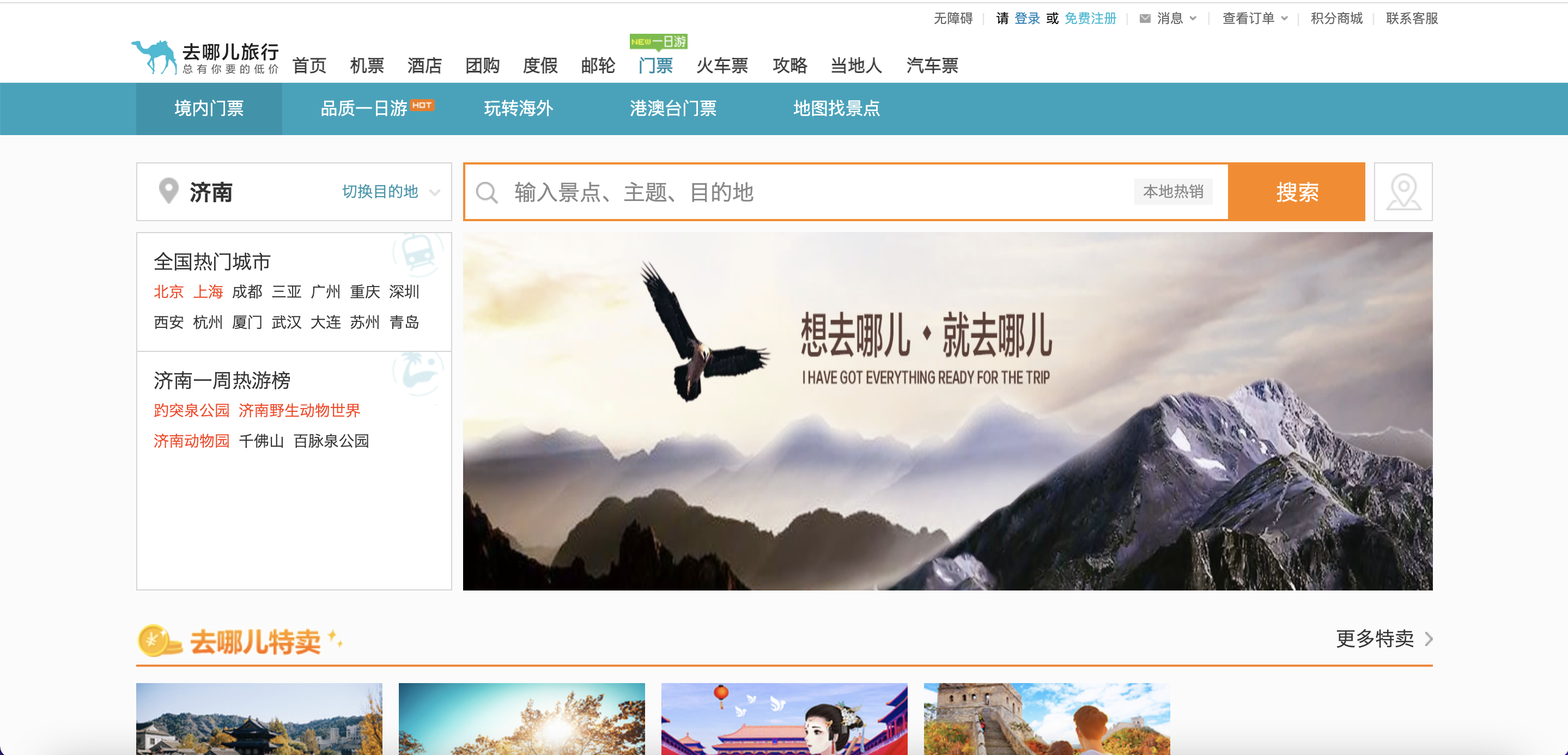
3、假期出游建议

4、景区价格分析

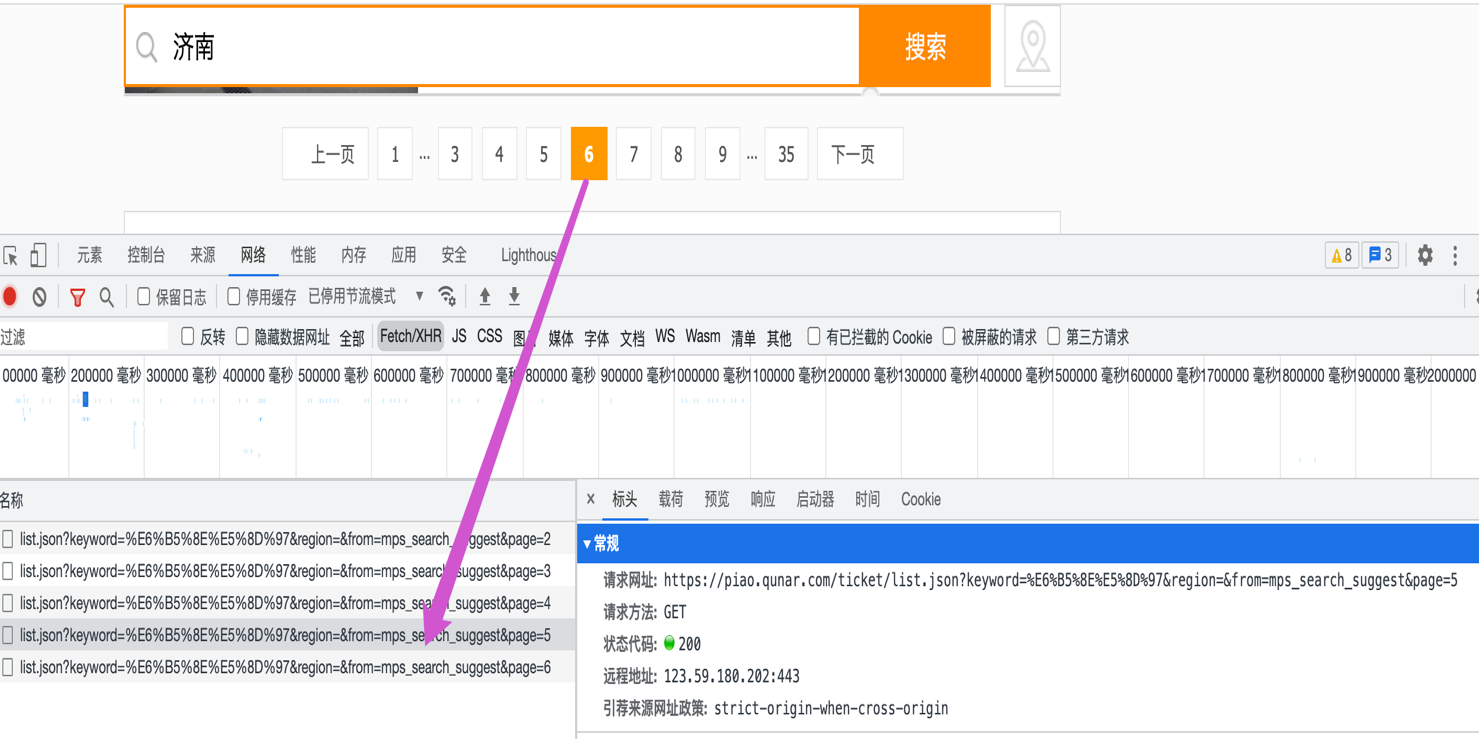
1. **数据获取方法及过程（附主要代码（3-5行））**

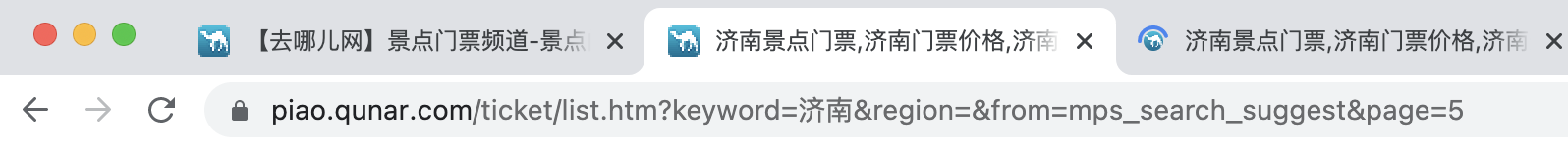
主要方法：request+json实现网页数据爬取

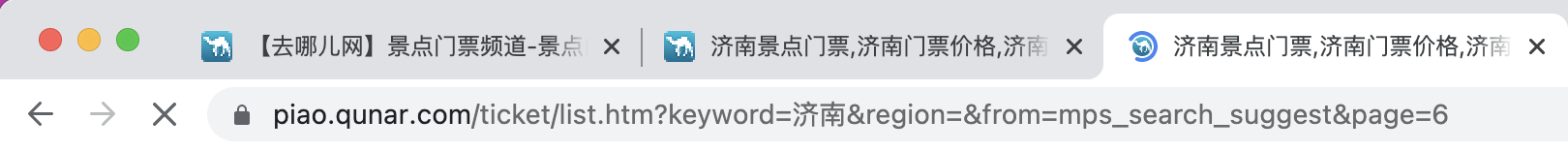
1.分析网页



1.1 F12打开浏览器调试窗口，查找加载数据的url，点击每一页观察网址变化：

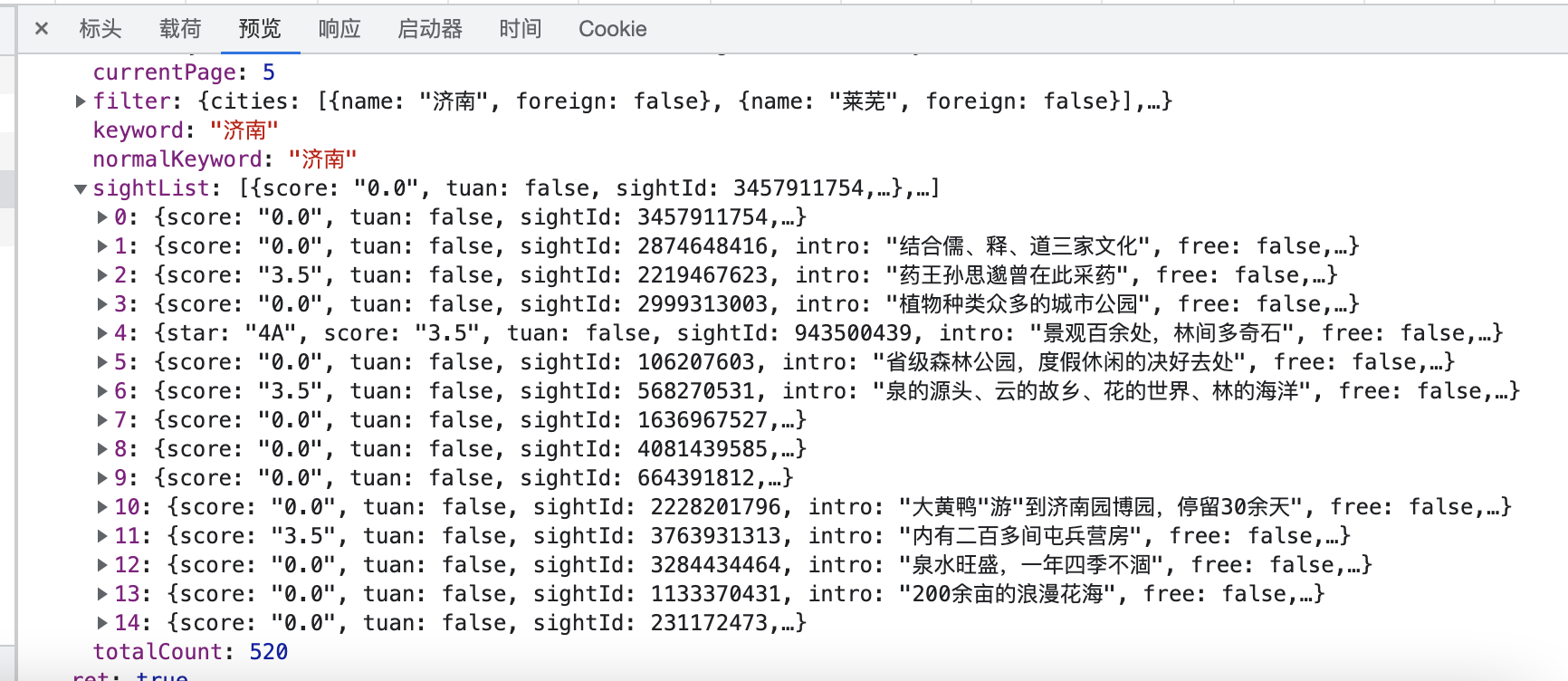






可以看到，每页网址只有page的值在发生变化，keyword参数可以直接写中文。

看一下接口返回信息：发现返回的是json字串



2. 爬取数据

2.1 导入模块包 安装缺失的包

2.2 爬取每个行政区数据

def get\_city\_scenic(city, page):

ua = UserAgent(verify\_ssl=False)

headers = {'User-Agent': ua.random}

url= f'https://piao.qunar.com/ticket/list.json?keyword={city}&region=&from=mps\_search\_suggest&sort=pp&page={page}'

result = requests.get(url, headers=headers, timeout=10)

result.raise\_for\_status()

return get\_scenic\_info(city, result.text)

2.3 爬取每一页的数据（设置所需字段）

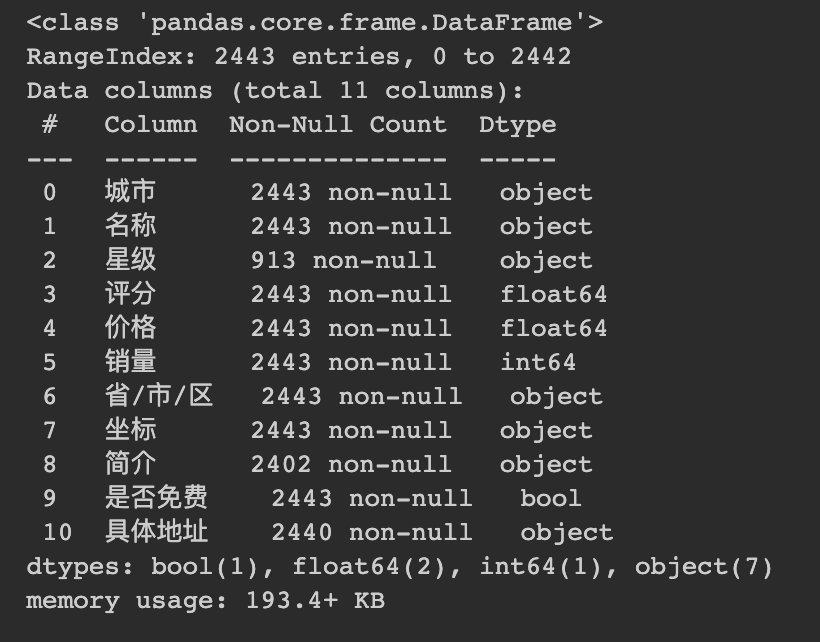
2.4 循环爬取每个行政区的每一页数据

2.5 保存数据，用openpyxl将数据保存到Excel中

1. **数据探索（根据需要加数据预处理、数据可视化）**

使用Pandas数据处理

1. 读取数据
2. 查看索引、数据类型和内存信息



1. 查看数值型列汇总统计



1. 查看销量为0的行



1. 去除销量为0的行数据
2. 统计各列空值



1. 将星级缺失值用‘未知’填充
2. 也可以将所有缺失值都用‘未知’填充



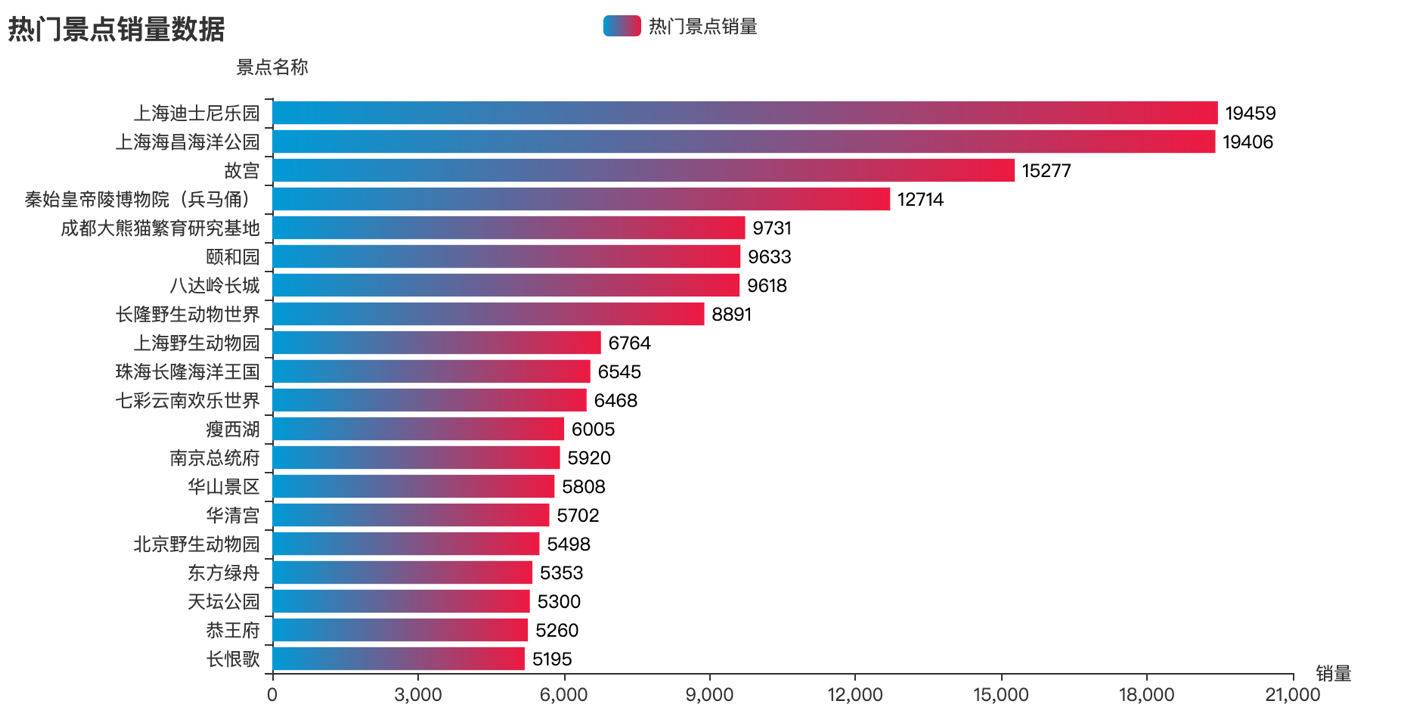
1. 按销量排序



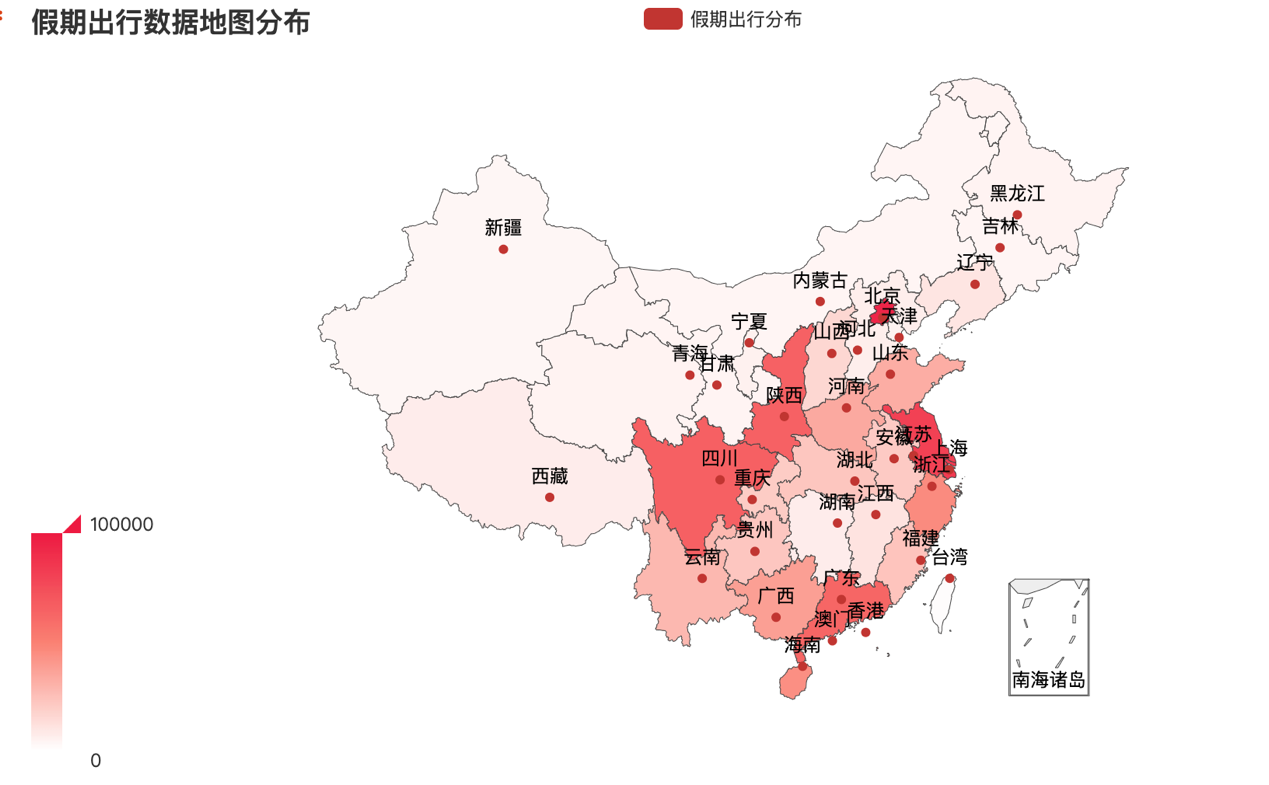
1. **数据可视化**

使用Pyecharts数据可视化

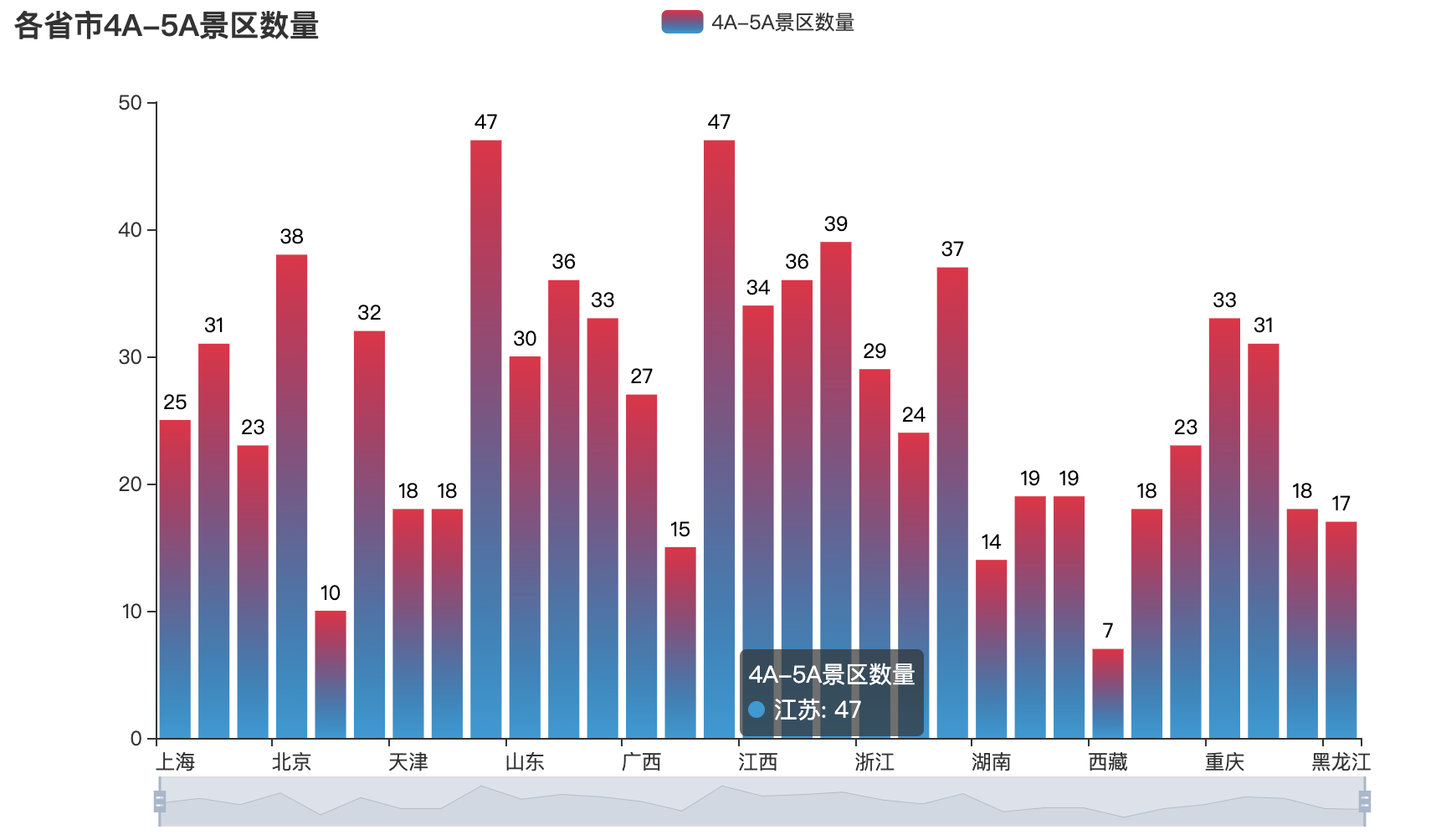
1. 销量前20热门景点数据



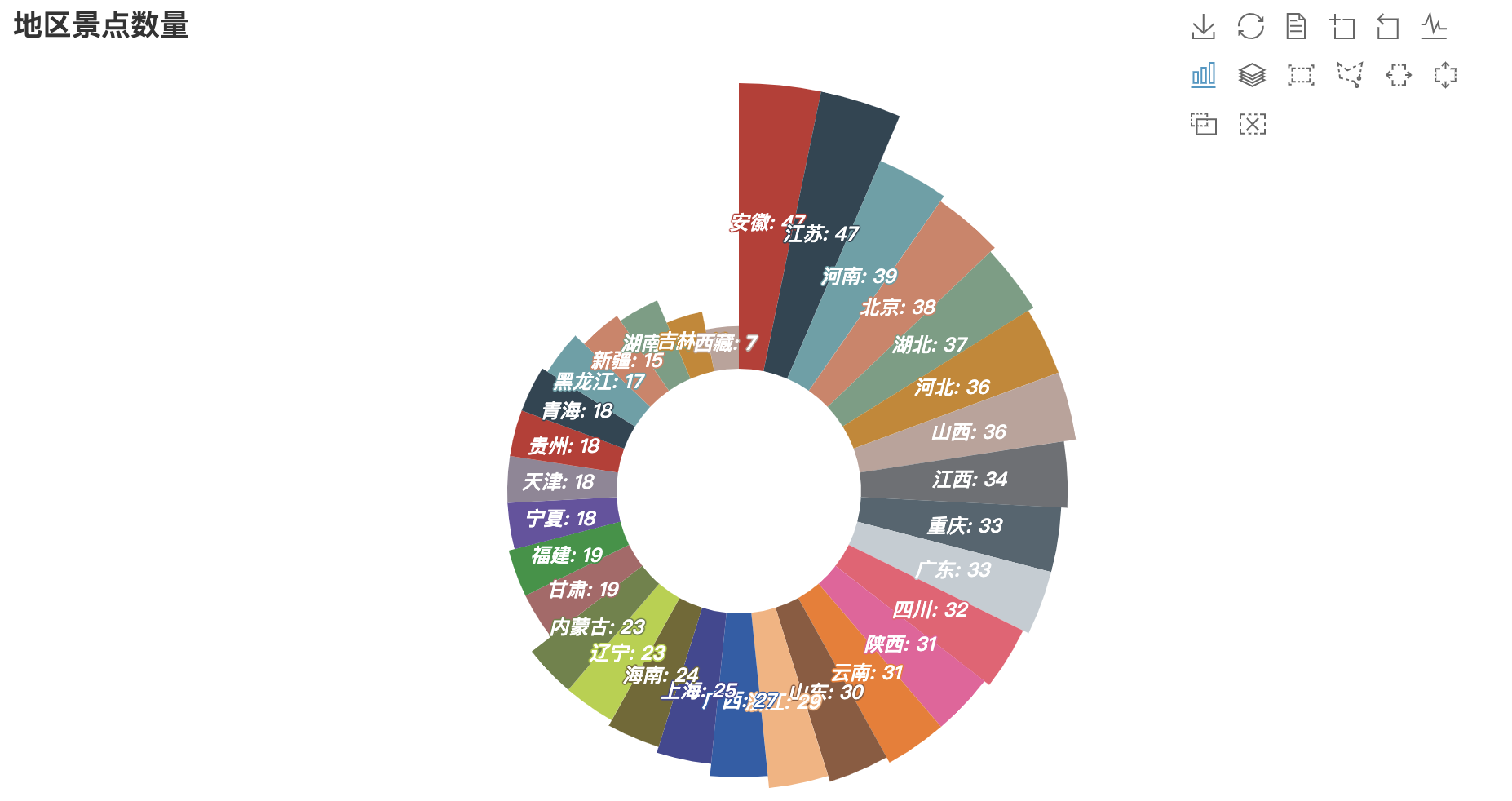
1. 假期出行数据全国地图分布



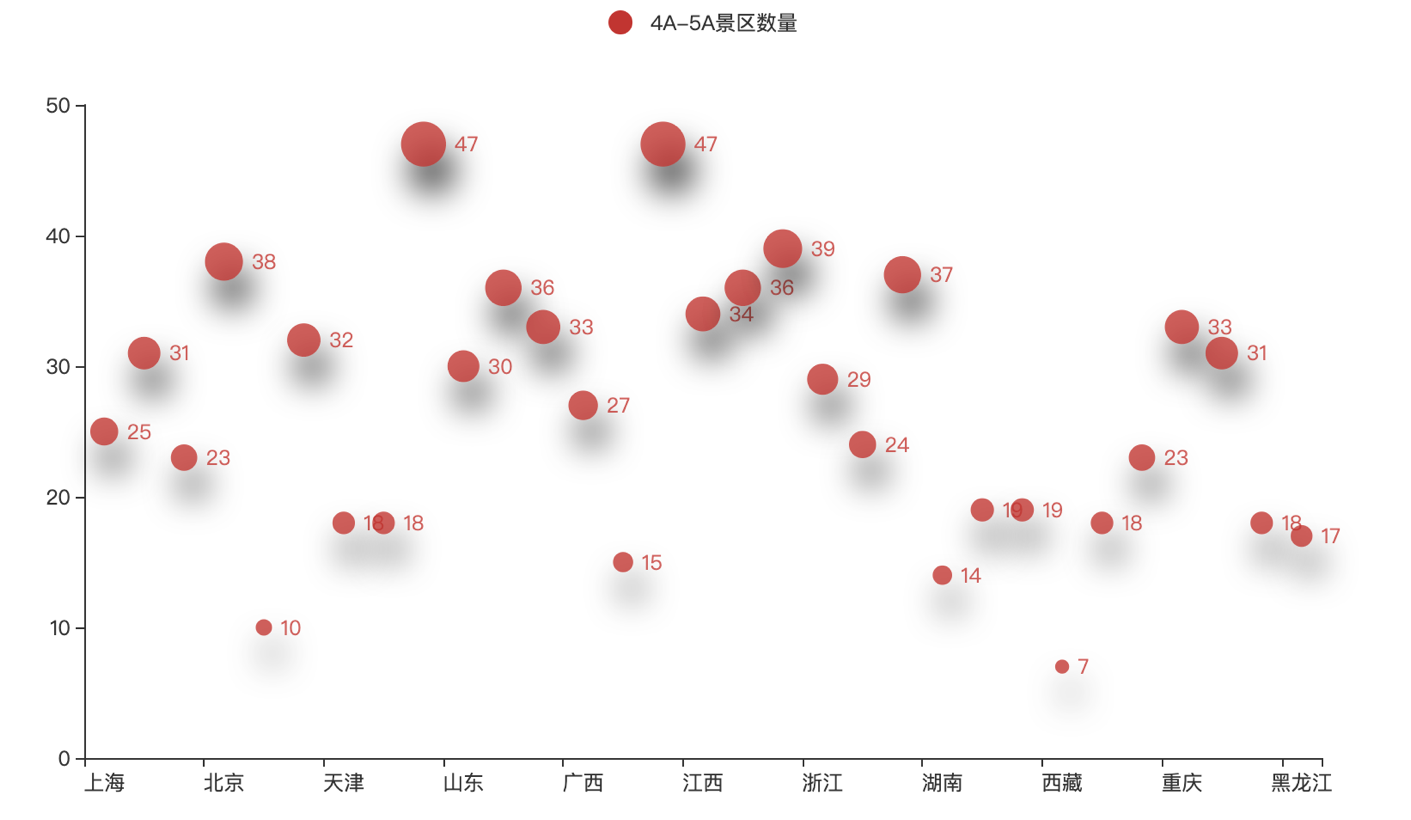
1. 各省市4A-5A景区数量柱状图



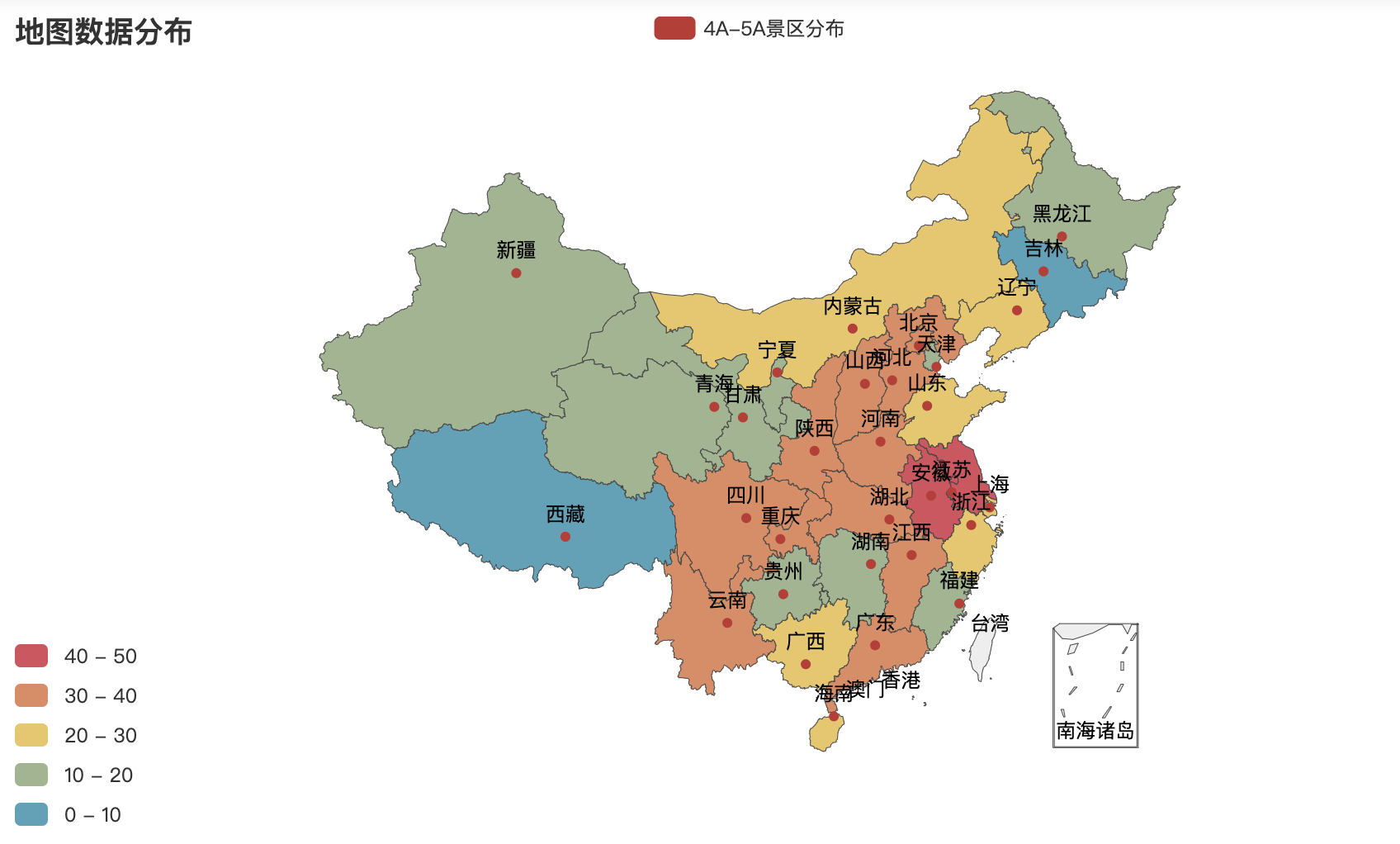
1. 各省市4A-5A景区数量玫瑰图

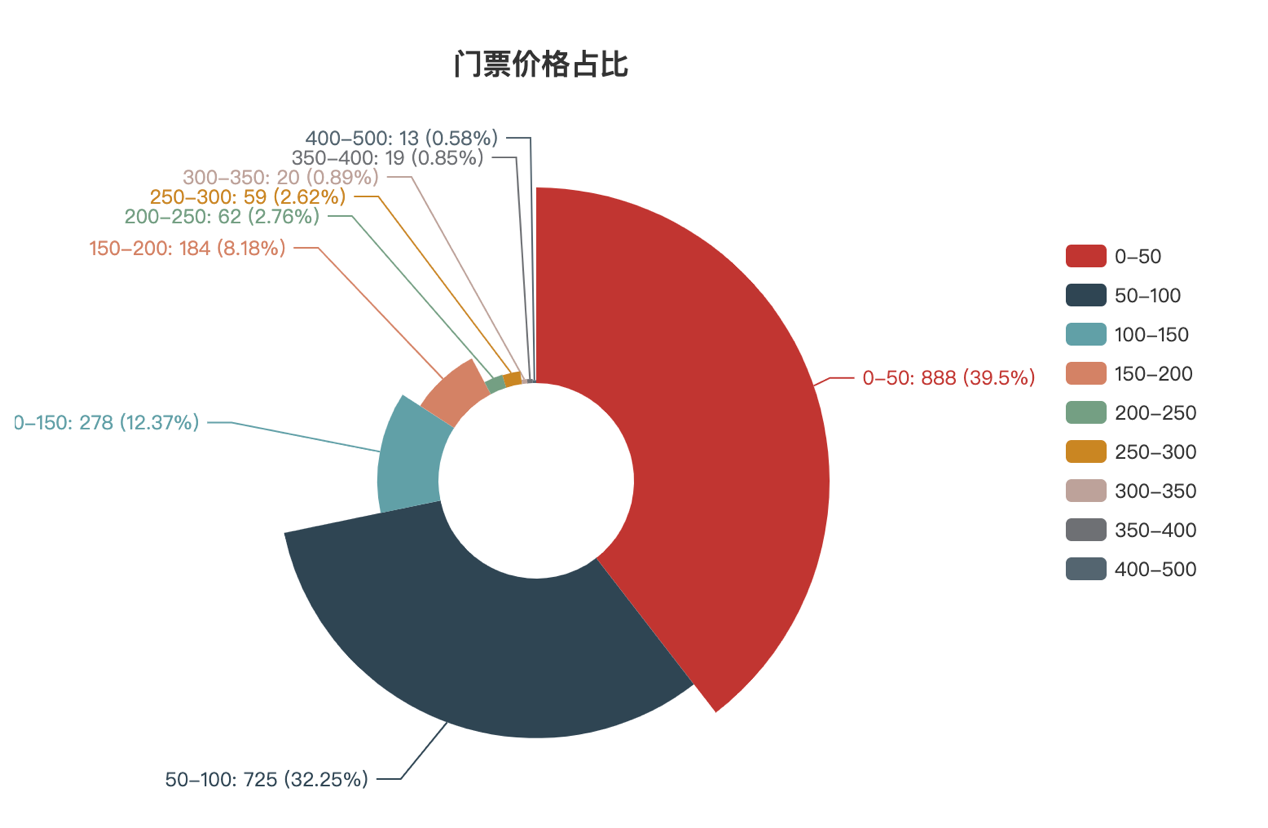


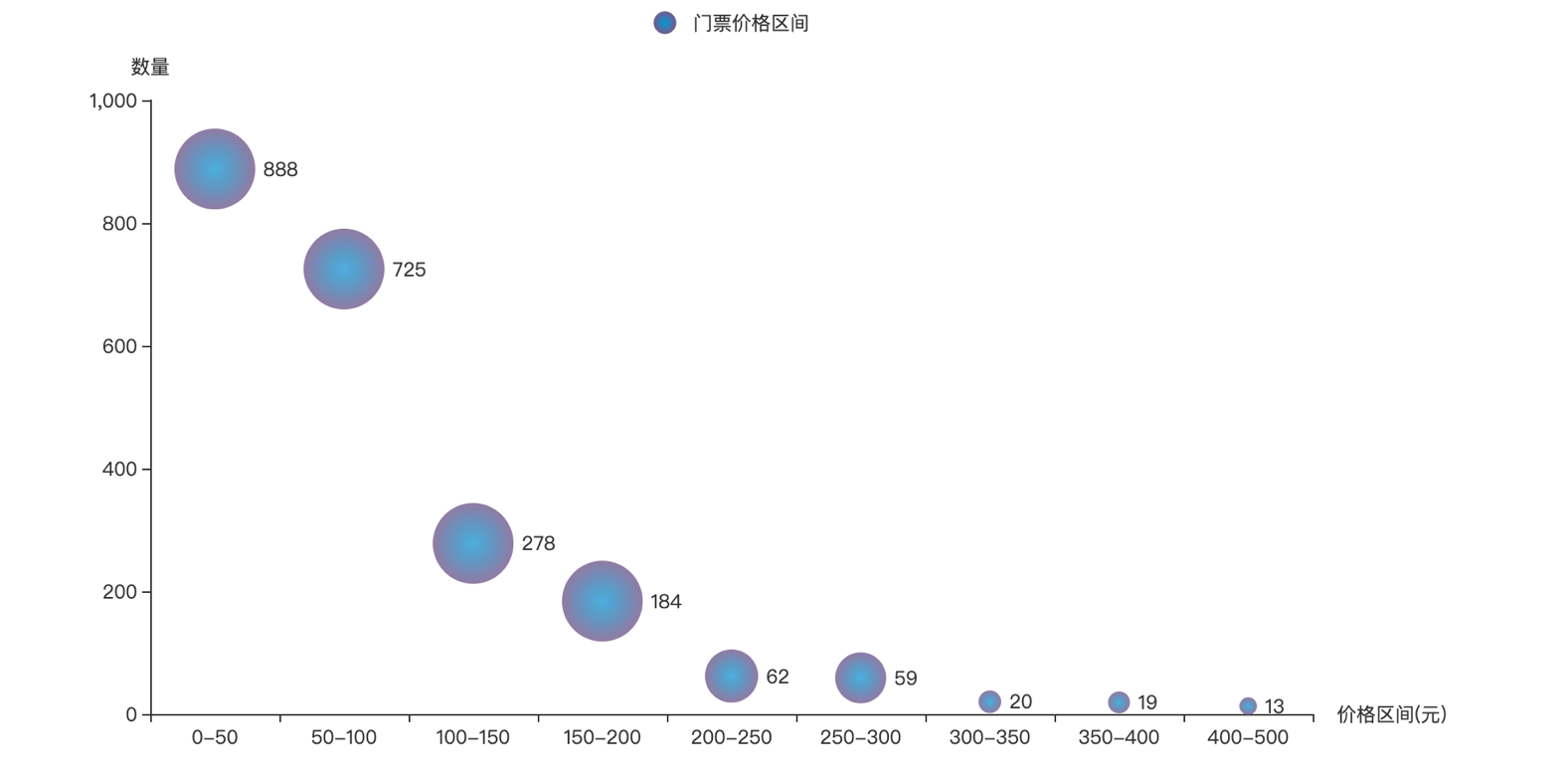
1. 各省市4A-5A景区数量阴影散点图



1. 各省市4A-5A景区地图分布

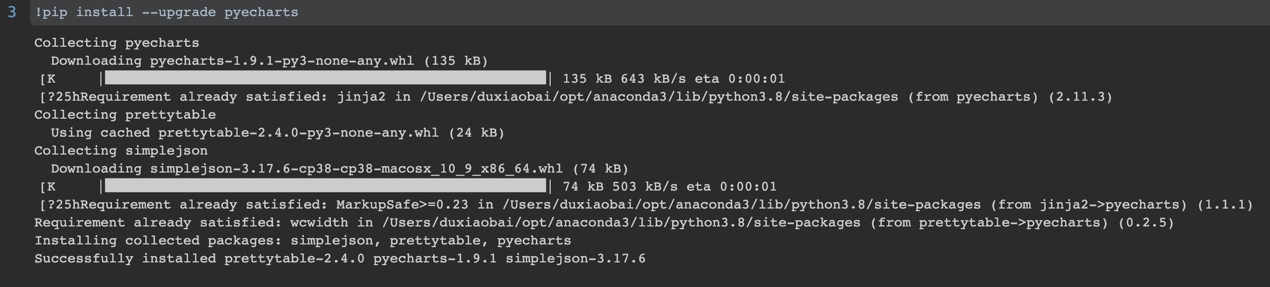
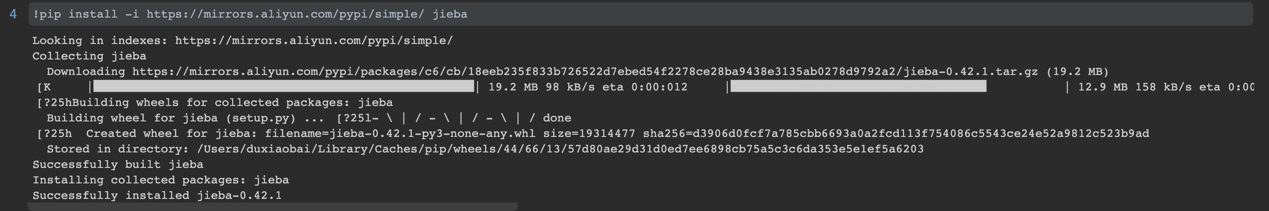


1. 门票价格区间占比玫瑰图
2. 门票价格区间数量散点图



1. 景点简介词云
2. **结论**
   * 1. 销量为0的数据一共有123行
     2. 去除后销量为0的数据后还剩下2320行数据
     3. 星级存在1407个空值，简介37个空值，具体地址2个空值，其他列不存在空值，数据还算比较完整
     4. 华东、华南、华中等地区属于国民出游热点地区，尤其是北京、上海、江苏、广东、四川、陕西等地区出行比较密集
     5. 江苏、安徽、河南、北京、湖北等地区4A、5A级景区数量比较多
     6. 门票价格100以内居多，大概占比70%，还是比较实惠的，而且一般景区还存在学生优惠等
3. **遇到问题及解决办法**

1.导包出现错误，在Pycharm中直接导包失败，后续改为直接pip方式完成包的导入



2. 地图显示部分需要用到 pyecharts，安装过旧版本需要升级，直接直接pip安装就是最新版本

3.没有区段标签，实际上很难区玫瑰图图段的大小，因此确保图表的所有区域都清楚地标记。 此外值得考虑的是所使用的类别的数量; 太多不同的段使得每个段之间难以区分

4.数据比较多比较迷惑。 颜色和数据范围如何正确的应用到图表中，是一个看起来比较简单却又有点难度的问题和细节，容易造成显示混乱。

1. **收获**

前面学了很多和数据挖掘分析Python的基础知识，可以简单的实现代码了，但是使用上还很不熟练，做不到如臂使指。这就要求更加努力去夯实基础，将自己的基础打牢。通过理解数据、数据清洗、数据可视化并合理使用图表。要解决的问题貌似很简单，但是真正在处理每一步的过程中，总是会出现一些错误，要想所有知识和技能要想全面掌握的话，还是得多加练习才可以。