1.1 利用爬虫技术获取深圳市数据分析师岗位数据

利用 Python 语言和 Drissionpage 库来模拟浏览器行为并访问目标网站，爬取了 BOSS 直聘网站上的深圳市数据分析师岗位的十页数据，一共 301 条数据，原始数据的例子请参考图 1.

公司	公司规模	职位	地区	地点	学历	薪资	技能要求
星普云	100-499人	高级、资	福田区	会展中心	大专	15-20K	数学/统计相关专业数据分析/挖掘经验计算机相关专业Finereport帆软报表开发
美图公司	1000-9999	数据分析	南山区	科技园	学历不限	15-25K·1	数据分析/挖掘经验SQLPython数学/统计相关专业
腾讯	10000人以	腾讯地图	南山区	科技园	本科	30-50K·1	数学/统计相关专业数据分析/挖掘经验机器学习建模经验
影石Inste	1000-9999	数据分析	宝安区	宝安中心[	本科	15-22K·1	PythonSQLTableau数学/统计相关专业计算机相关专业数据分析/挖掘经验
SUTPC	1000-9999	数据分析	福田区	车公庙	本科	15-20K·1	Python数学/统计相关专业数据分析/挖掘经验机器学习建模经验
点宽科技	20-99人	数据分析	南山区	科技园	硕士	12-18K	Python数学/统计相关专业金融行业数据相关经验数据分析/挖掘经验
康益药食[	0-20人	数据分析	龙华区	民治	本科	10-11K	
深圳前海[	10000人以	数据分析	盐田区	沙头角	本科	13-17K	SASSPSSR
华为	10000人以	数据分析	龙岗区	坂田	本科	30-60K	数据分析数据处理数据仓库
神策数据	1000-9999	数据分析	南山区	科技园	本科	15-20K	电商/零售行业数据相关经验文娱/游戏行业数据相关经验数据分析银行金融
样样进出	0-20人	数据分析	龙岗区	大运	本科	5-9K	数据分析电子商务用户研究
诚则成大	100-499人	数据分析	龙岗区	坂田	本科	8-10K	Excel相关分析数据分析/挖掘经验描述统计国内院校优先多元统计分析SPSS数学/统计相关专业
中银信息	100-499人	数据分析	罗湖区	莲塘	本科	15-30K	Python
SenseTime	1000-9999	数据分析	南山区	科技园	本科	15-30K·1	人工智能PythonExcel
腾讯	10000人以	数据分析	南山区	科技园	本科	10-15K	SQLPython数学/统计相关专业计算机相关专业数据分析/挖掘经验信息流内容行业短视频
云博智联	20-99人	数据分析	南山区	南山京基[	本科	7-8K	数据分析SPSSStata大数据
塔思科技	20-99人	数据分析	宝安区	新安	本科	16-26K·1	SQLPythonHive游戏数据分析
慧宝居	20-99人	数据分析	龙岗区	坂田	大专	8-10K	数学/统计相关专业数据分析/挖掘经验电商/零售行业数据相关经验数据分析师数据汇总电商数据分析师
前海梧桐	20-99人	数据分析	南山区	科技园	硕士	15-30K	JavaSQLPythonRExcel
上海从鲸[	500-999人	数据分析	南山区	科技园	本科	30-50K·1	SQLHive数学/统计相关专业计算机相关专业数据分析/挖掘经验机器学习建模经验
第五空间	10000人以	数据分析	福田区	梅林	本科	15-20K·1	SQLPython

图1 爬取深圳市数据分析师岗位数据（节选）

使用 Python 软件对网页数据提前进行筛选处理，主要收集了公司、公司规模、职位、地区、地点、学历、薪资和技能要求的数据，并将文件保存为 xlsx 后缀文件。

1.2 利用 Matplotlib 技术绘制深圳市数据分析师岗位分布情况饼图

利用 Pandas 库读取之前通过网络爬虫技术收集并保存至 Excel 文件中的数据集。然后，针对“地区”这一字段进行了频数统计，以获取每个区域内数据分析师岗位的数量。为了直观展示深圳市数据分析师岗位在不同区域内的分布情况，我利用 Matplotlib 库绘制了地区分布饼图作为主要的数据可视化工具。饼图能够清晰地反映出每个区域内职位数量占总数的比例，快速了解数据分析师职位在深圳各区域的集中度。地区分布饼图请参考图 2

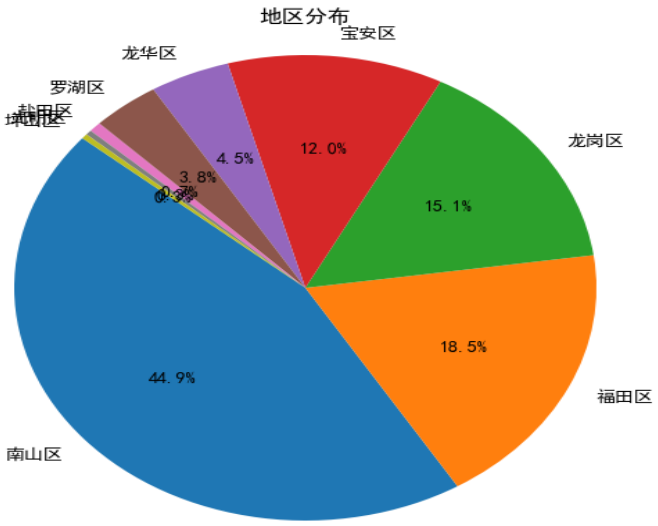

图2 地区分布饼图

根据饼图显示，南山区占比最高，占比为 44.9%，福田区次之，占比为 18.5%，这表明南山区和福田区是深圳市数据分析师岗位较为集中的区域，可能是因为南山区和福田区作为深圳市的经济中心，拥有大量的高新技术企业和金融机构，在电子信息、生物医药、新能源及新材料等领域有深厚积累，孕育了华为、腾讯、比亚迪等世界级巨头，以及大疆、柔宇科技等独角兽企业；在金融领域中，深圳金融行业上市公司与总市值均居全国前列，招商银行、平安保险等金融巨头与新兴金融科技企业交相辉映，形成了以福田区为核心的金融高地。此外，深圳积极推动企业加强与国内外大专院校科研院所的合作，初步建立起了适合深圳特点的 以市场为导向、企业为主体、内地高校和科研院所为依托、国外研究开发机构为补充的技术开发体系，[3]众多高校研究院坐落于南山区，例如清华大学研究院、华中科技大学研究院、南京大学研究院等高校以及深圳大学、深圳理工大学等综合性大学，促使南山区拥有丰富的人才储备以及科技创新实力，进一步促进了该区域对数据分析师的需求。

龙岗区是第三大集中区域，占比为 15.1%，宝安区暂列第四，占比为 12.0%，这说明了龙岗区和宝安区数据分析师需求增长潜力大，这可能与深圳市政策支持和产业布局有关，宝安区正在推进“2024 年宝安区高层次科技创新人才及团队资助项目”，旨在支持科技创新人才和团队的发展，龙岗区对高新技术企业进行认定奖励，首次申请给予 10 万元，通过重新认定的，每次给予 5 万元奖励，促进企业入驻。

龙华区、罗湖区和盐田区的数据分析师岗位占比分别为 4.5%、3.8%和 3.0%，合计约占总体的 11.3%。虽然这些区域的职位比例相对较小，但它们各自具有独特的经济特征和发展潜力，罗湖区出台了扶持软件信息和人工智能产业发展相关政策，盐田区加大招商引资力度，龙华区出台了激励企业研发投入相关条例，有望在未来进一步提升这些区域在数据分析领域的重要性。

1.3 利用 Matplotlib 技术绘制薪资直方图

利用 Pandas 库读取之前通过网络爬虫技术收集的数据，使用 Matplotlib 库对薪资数据进行分组，确定薪资区间，绘制了薪资分布直方图，直方图的横轴表示薪资区间，纵轴表示每个薪资区间内的岗位数量。薪资分布直方图参考图 3.

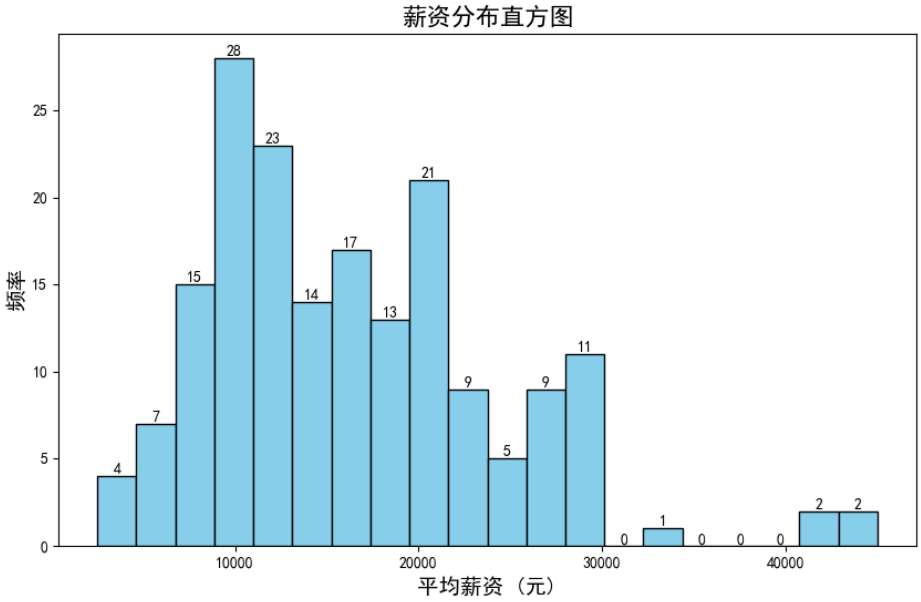

图3 薪资分布直方图

从薪资直方图来看，工资主要集中在 10000 元到 20000 元之间。10000 元到 12000 元区间岗位数量最多，为 28 个，表明市场上有大量的初级和中级数据分析师岗位需求，高薪资区间（20000 以上）岗位数量较少，但仍然有一定的需求，岗位主要分布在华为、腾讯、字节跳动、vivo、荣耀等互联网中上游企业，以及影石等独角兽企业和 SHEIN 等电商行业，侧面反映市场对高级数据分析师的需求。

1.4 利用 Wordcloud 库绘制技能要求词云图

使用 Pandas 库读取相关数据，提取技能要求关键词，并进行词频统计，使用 Wordcloud 生成词云图，词云图中词的大小代表词频，词频越高，词在图中的字体越大。要求词云图参考图 4

图4 技能要求词云图

根据技能要求词云图，从技能方面，统计相关专业是祠云图中最大的词，表明统计学相关专业背景是数据分析师岗位的基本要求，数据分析技能也是数据分析师的核心技能，同时企业重视数据挖掘能力。从工具方面，数据分析师应该熟练使用 SQL 和 Python 语言。与此同时，BI 相关技能需求较高，Excel 作为基础工具，数据可视化工具 Tableau 需求也较为明

显。从行业相关技能来看，零售、电商和金融行业对数据分析师需求较高，重视数据分析能力，这可能是因为数据分析对数据驱动决策有紧密联系。

爬虫代码展示

```
    from DrissionPage import ChromiumPage #导入用于自动化浏览器操作的库
    from pprint import pprint
    import json #用于处理 json 数据
    import pandas as pd #用于数据分析和处理
    import os #用于文件路径操作
    dp = ChromiumPage() #初始化 Chromium 浏览器实例
    dp.listen.start('zhipin.com/wapi/zpgeek/search/joblist.json?')#设置监听器以捕获特定 URL 模式的请求响应
    all_data=[]
    dp.get('https://www.zhipin.com/web/geek/job?query=%E6%95%B0%E6%8D%AE%E5%88%86%E6%9E%90%E5%B8%88&city=101280600')#打开目标网页
    for page in range(1,11):
        print(f'正在采集第{page}页数据')
        dp.scroll.to_bottom()#滚动页面到底部
        resp = dp.listen.wait()#等待监听器捕获到对应的 API 请求并获取响应内容
        json_data = resp.response.body#将响应体解析为 json 格式
        jobList = json_data['zpData']['jobList']#提取职位列表数据
        data = []
        for index in jobList:
            dit = {
                '公司': index.get('brandName'),
                '公司规模': index.get('brandScaleName'),
                '职位': index.get('jobName'),
                '地区': index.get('areaDistrict'),
                '地点': index.get('businessDistrict'),
                '学历': index.get('jobDegree'),
                '薪资': index.get('salaryDesc'),
                '技能要求': ''.join(index.get('skills', []))
                }
            data.append(dit)#将每个职位信息添加到临时列表中
        all_data.extend(data)#将当前页面所有职位信息追加到总数据列表中
        dp.ele('css:.options-pages a:last-of-type').click()#点击进入下一页
        df = pd.DataFrame(all_data)#数据收集完毕后，创建 DataFrame 对象并将数据写入 Excel 文件
        file_path = 'C:/Users/886/Desktop/python2.xlsx'
        df.to_excel(file_path, index=False)
        print(f"Data has been saved to {file_path}")
```