Python语言程序设计

大作业文档

|  |  |
| --- | --- |
| 项目名称： | 前程无忧python岗位信息爬取和分析 |
| 专业： | 软件工程 |
| 班级： |  |
| 学号： |  |
| 姓名： |  |

目录

[一、 项目概述 3](#_Toc4356)

[二、 项目环境需求 3](#_Toc12833)

[1. 开发环境 3](#_Toc25675)

[2. 运行环境 3](#_Toc2747)

[3. 是否需要联网 4](#_Toc10321)

[三、 项目设计 4](#_Toc1485)

[1. 数据库设计 4](#_Toc20427)

[2. 工程结构目录 5](#_Toc31469)

[四、 项目实现 5](#_Toc3299)

[1. 数据获取实现 5](#_Toc20702)

[2. 数据存储 6](#_Toc15242)

[3. 数据呈现 6](#_Toc11036)

[4. 数据分析 9](#_Toc4948)

[5. 创新功能及新技术实现 16](#_Toc9425)

## 项目概述

爬取2023年前程无忧网站上搜索关键字“python”,所得到的数据约1000多条（只有这么多），并对这些数据进行清洗，分析及可视化处理。包括不同学历要求下岗位数量分布，不同工作经验相对真实月薪，各地区相对真实月薪，各城市地区的岗位数量分布，福利待遇关键词，平均月薪，公司最喜欢招聘有多少年经验的人等7种分析。

## 项目环境需求

### 开发环境

电脑系统：Windows11

开发工具：Pycharm

数据库：MySQL，Navicat

数据采集工具：八爪鱼

### 运行环境

Python解释器：Python3.1.1

虚拟环境：Anaconda

Python包：pandas，matplotlib.pyplot，sqlalchemy，numpy，wordcloud，re

数据库：MySQL

### 是否需要联网

否

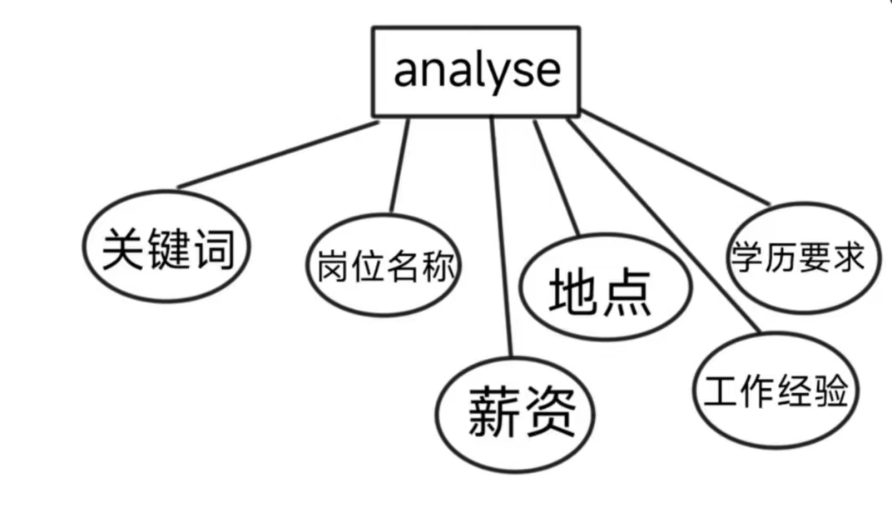
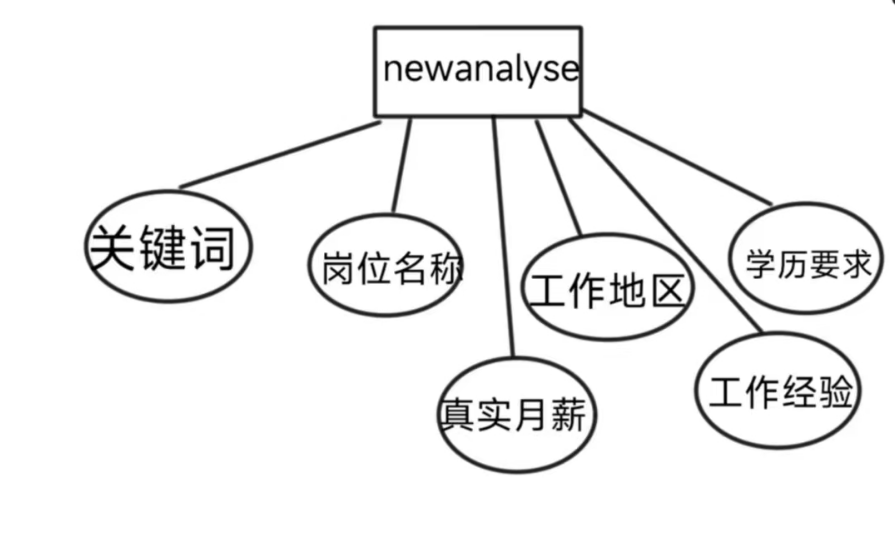
## 项目设计

### 数据库设计

1. 数据库设计概述

存储不同岗位的相关信息，包括关键词、岗位名称、真实月薪、工作地区、工作经验和学历要求等信息。

1. 逻辑结构设计（E-R图）



1. 物理结构设计数据表

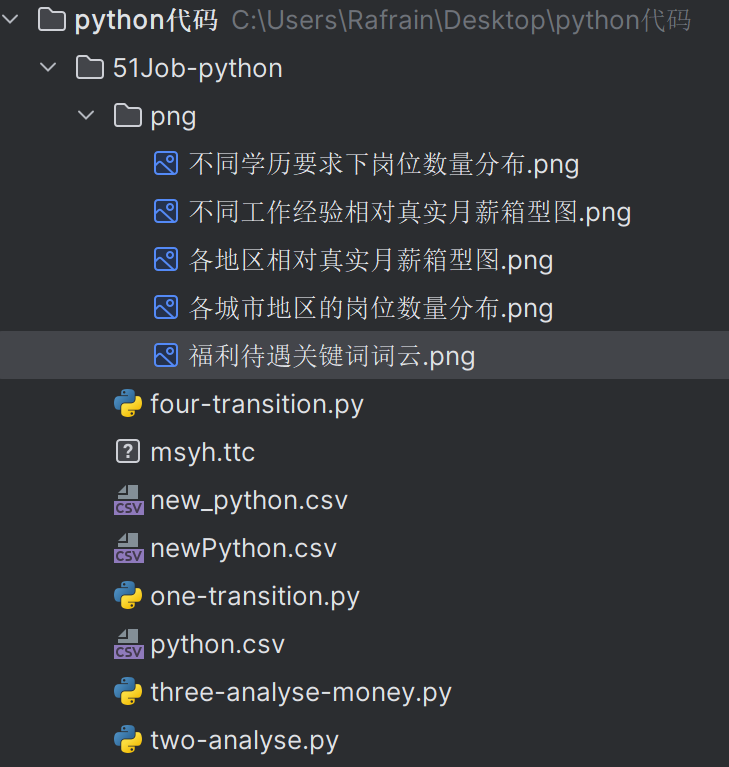
表1 analyse表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 空/非空 | 其他 | 字段简单描述 |
| 关键词 | Text | 非空 |  | HR画的最多的饼 |
| 岗位名称 | Text | 非空 |  |  |
| 薪资 | Text | 非空 |  | 格式不统一需要处理 |
| 地点 | Text | 非空 |  | 格式不统一需要处理 |
| 工作经验 | Text | 非空 |  |  |
| 学历 | Text | 非空 |  |  |

表2newanalyse表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 字段名 | 数据类型 | 空/非空 | 其他 | 字段简单描述 |
| 关键词 | Text | 非空 |  | HR画的最多的饼 |
| 岗位名称 | Text | 非空 |  |  |
| 真实月薪 | bigint | 非空 |  | 处理过的 |
| 工作地区 | Text | 非空 |  | 处理过的 |
| 工作经验 | Text | 非空 |  |  |
| 学历 | Text | 非空 |  |  |

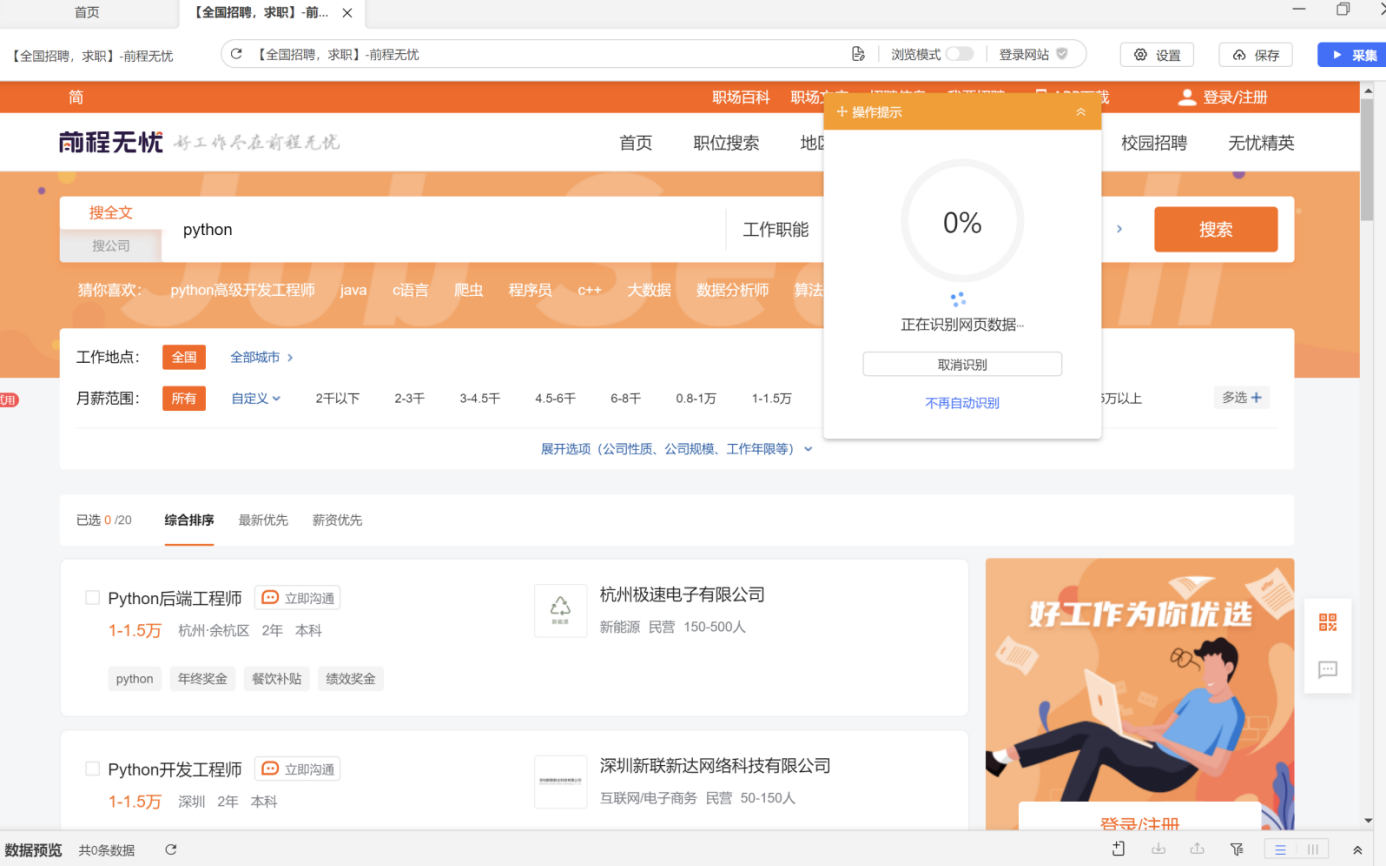
### 工程结构目录



## 项目实现

### 数据获取实现

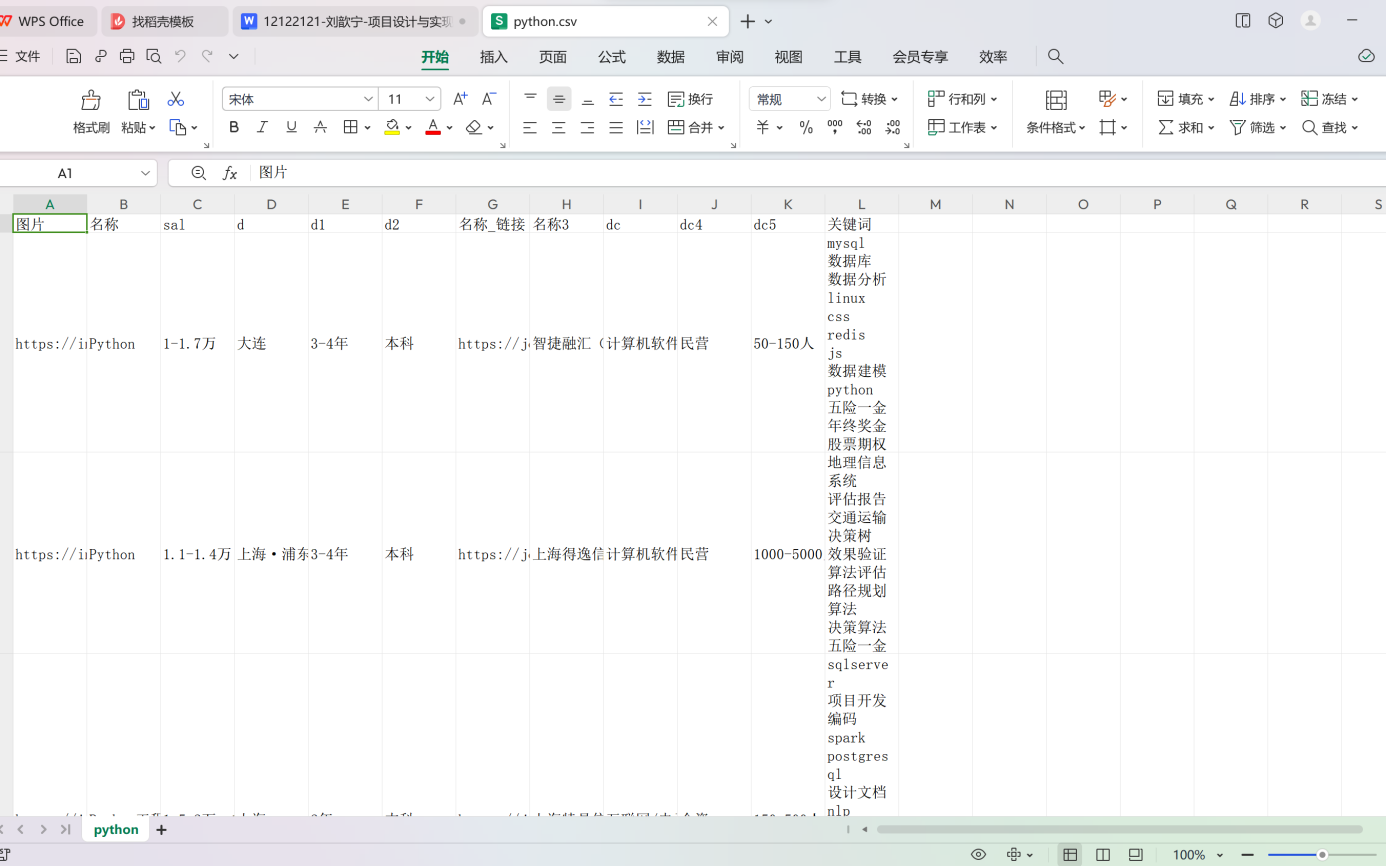
<https://we.51job.com/pc/search?keyword=python&searchType=2&sortType=0&metro=>

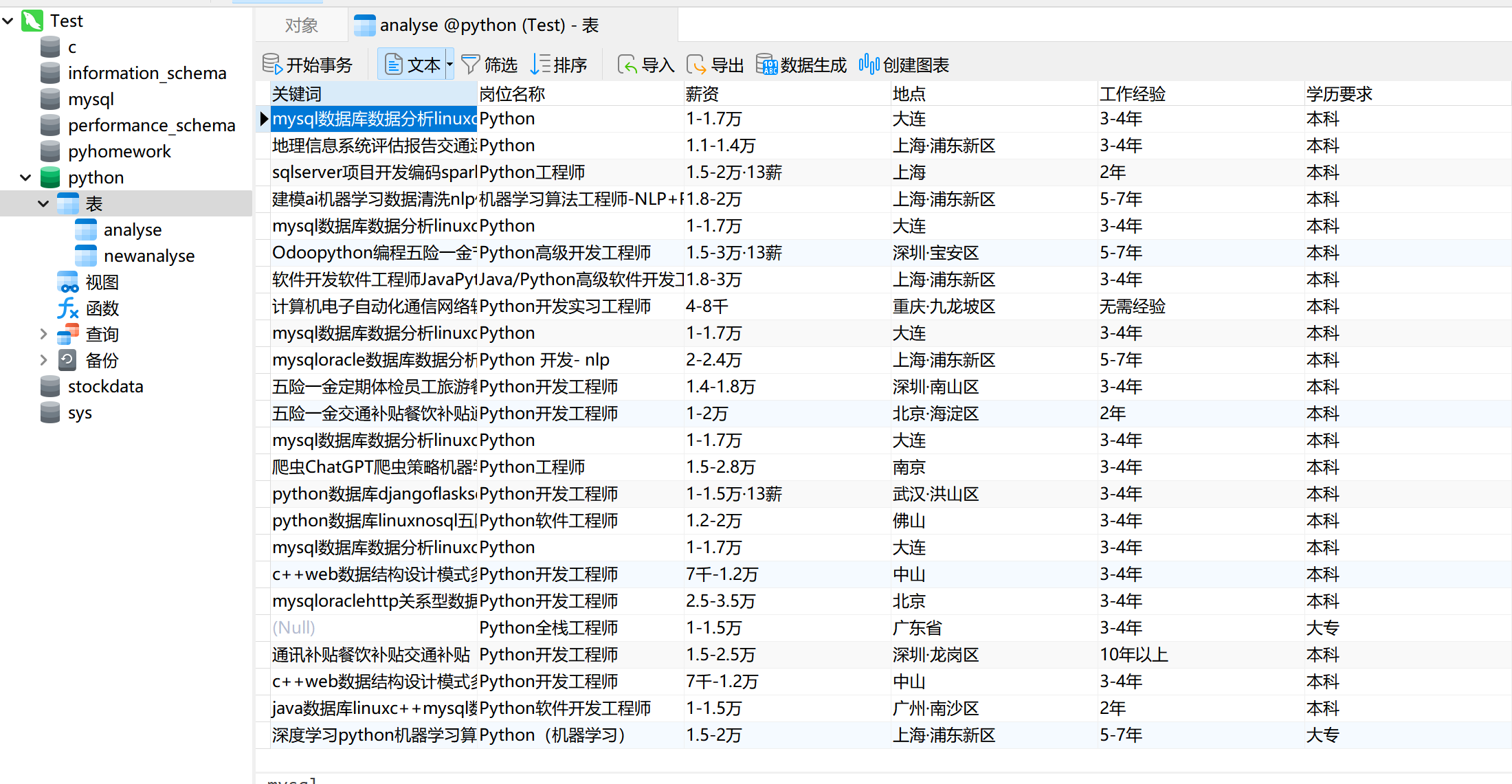


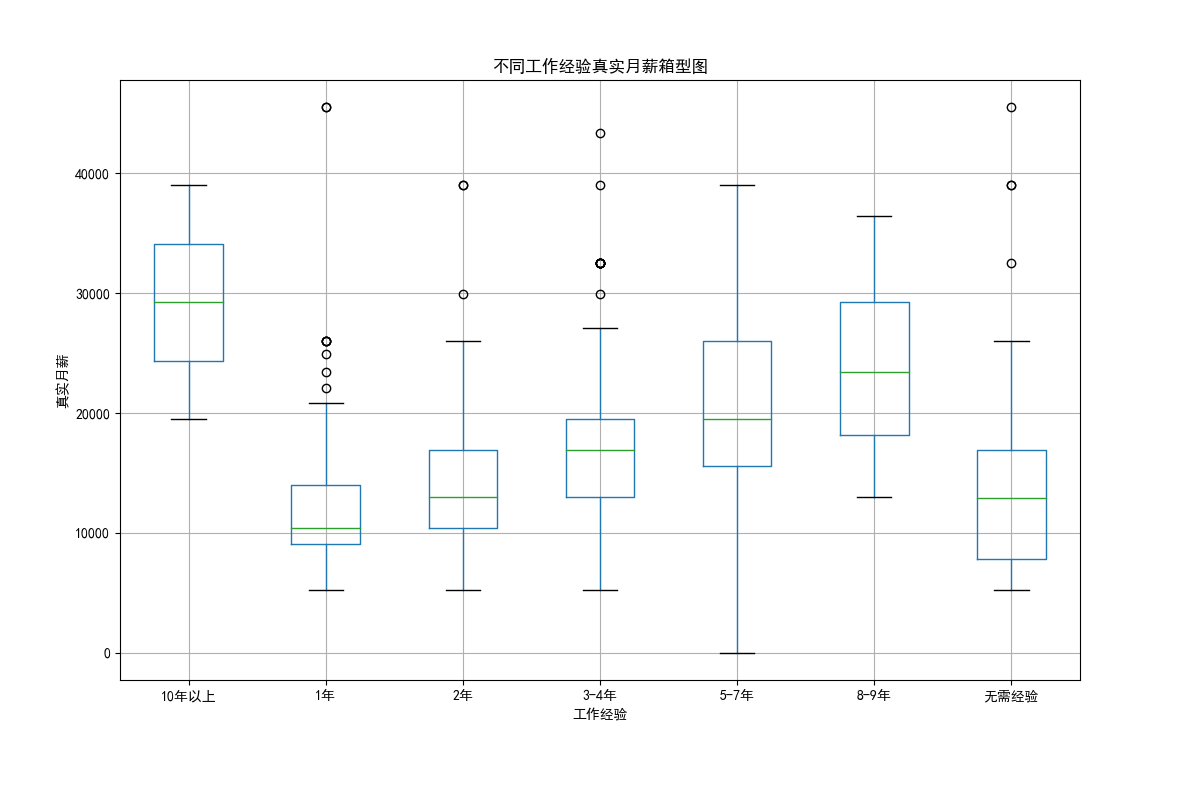
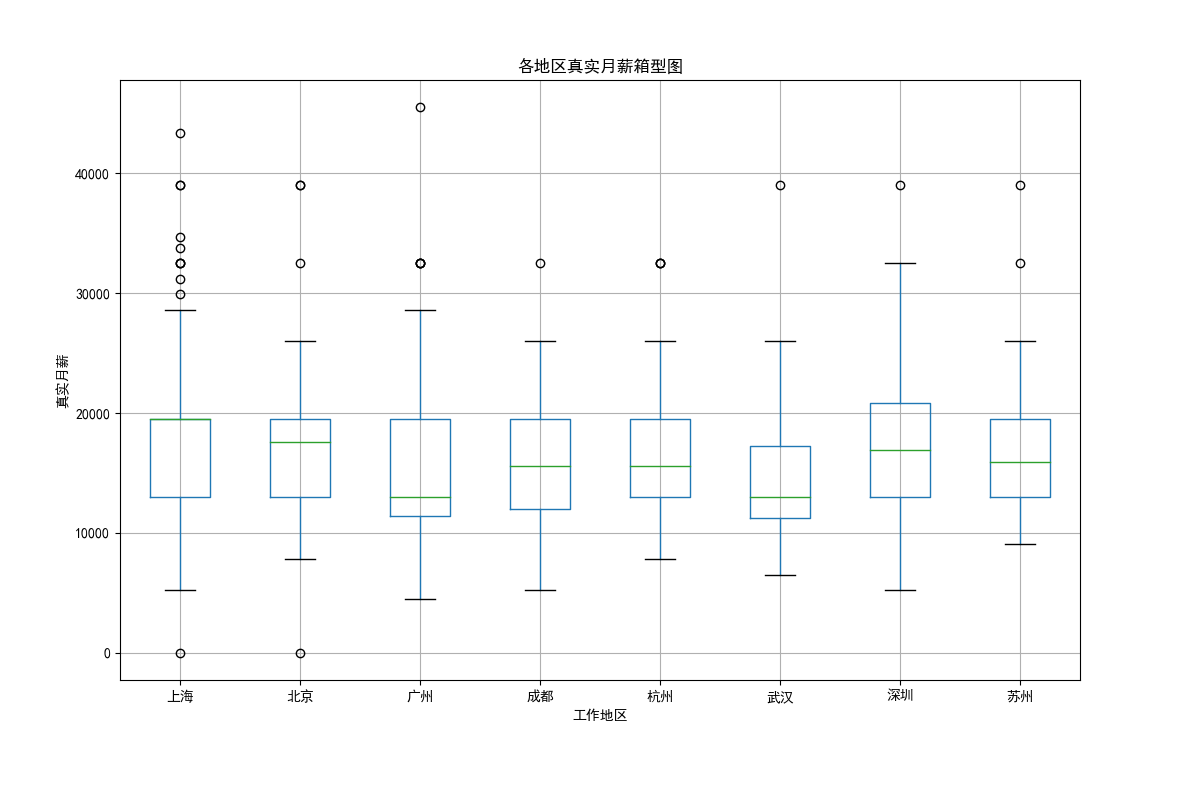
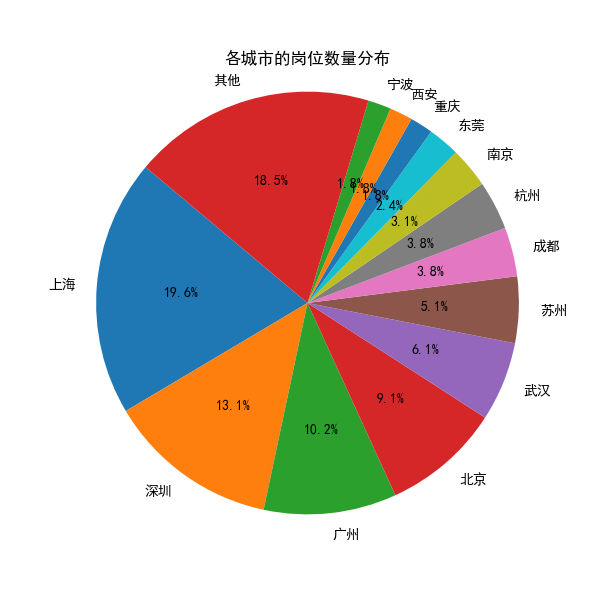
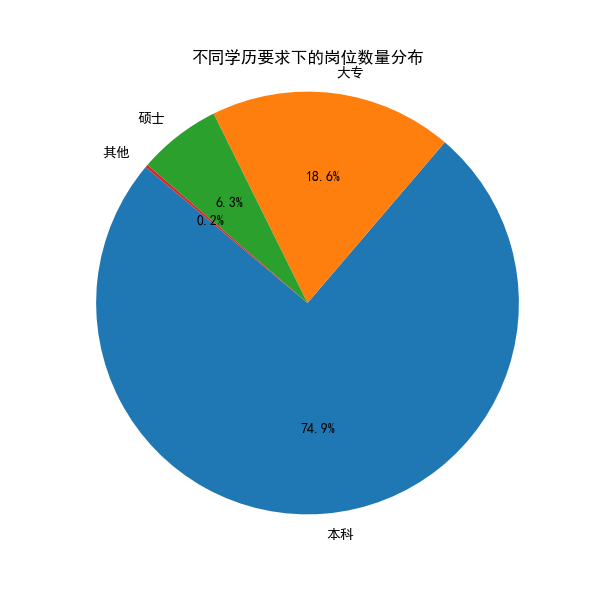
### 数据存储

使用该软件爬取数据后可自动保存为CSV文件

### 数据呈现（数据库及CSV文件、可视化）



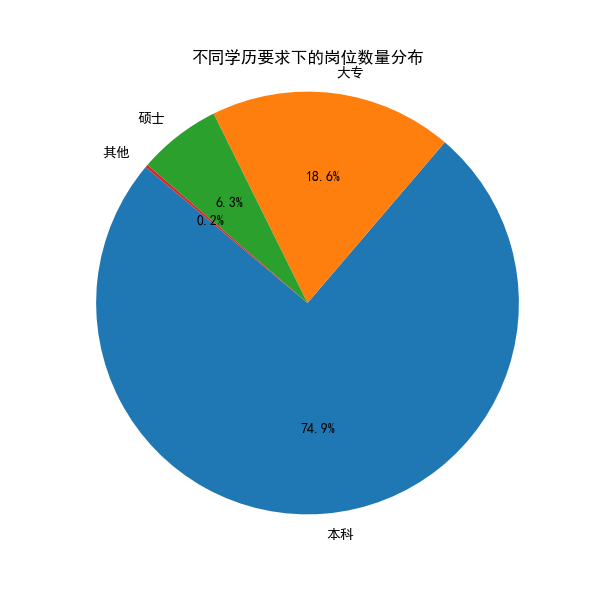
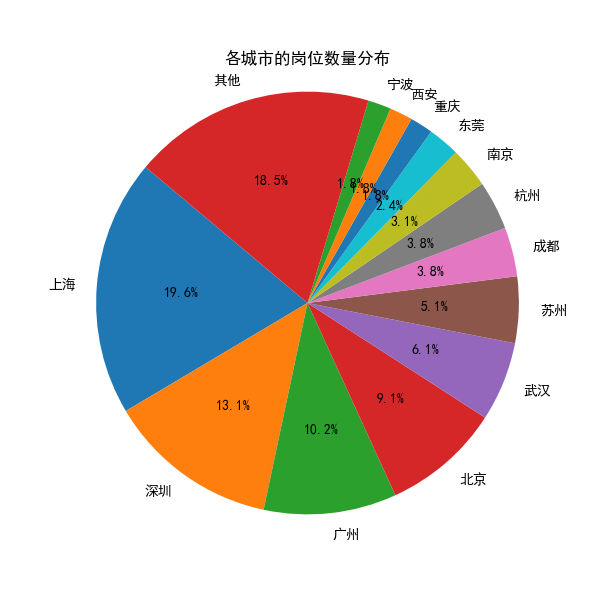




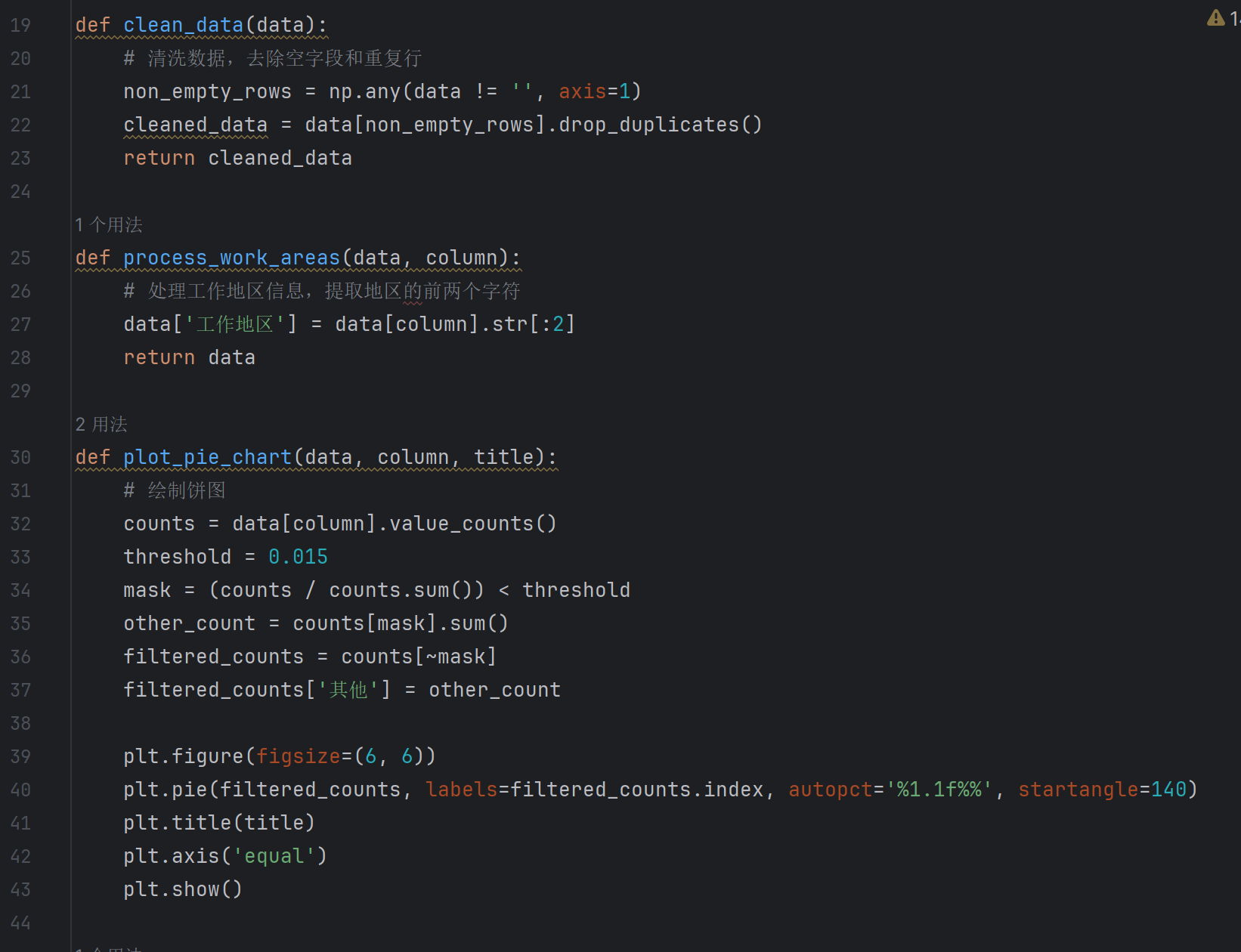
### 数据分析

1. 由于互联网行业是近20年跟着中国快速发展而高速发展的行业，所以在经济发展快的地区python相关岗位往往更多，在“各城市的岗位数量分布”饼图中我们可以很直观的发现，北上广深四个城市占据了中国内地超过50%的岗位，由此可以分析出经济发展好的地区python相关岗位更多这个结论，当应届生毕业之后选择去这些城市也会有更多的机会，发展也会相对来说好一些。

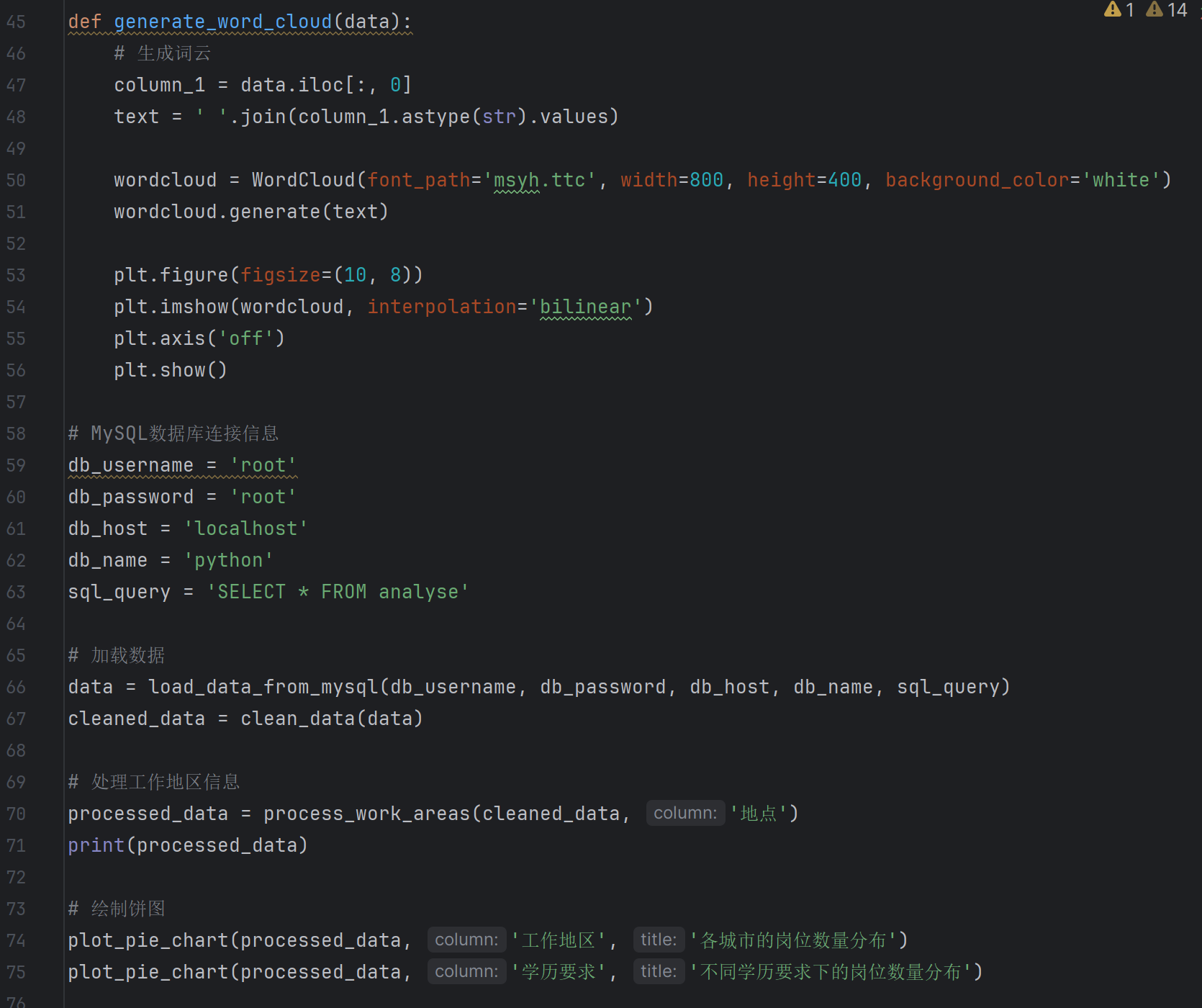
由于计算机岗位对员工技术的要求明显高于其他行业，所以大部分公司更需要有技术，经验（经验这点会在后面的数据分析中提到）的求职者，而不太特别看重学历，这点可以从“不同学历要求下的岗位数量分布”饼图中看到，本科学历占据了74.9%，大专学历占据18.6%，硕士学历只占6.3%，而博士学历只占不到0.2%。



**PS.这里附上关键代码**



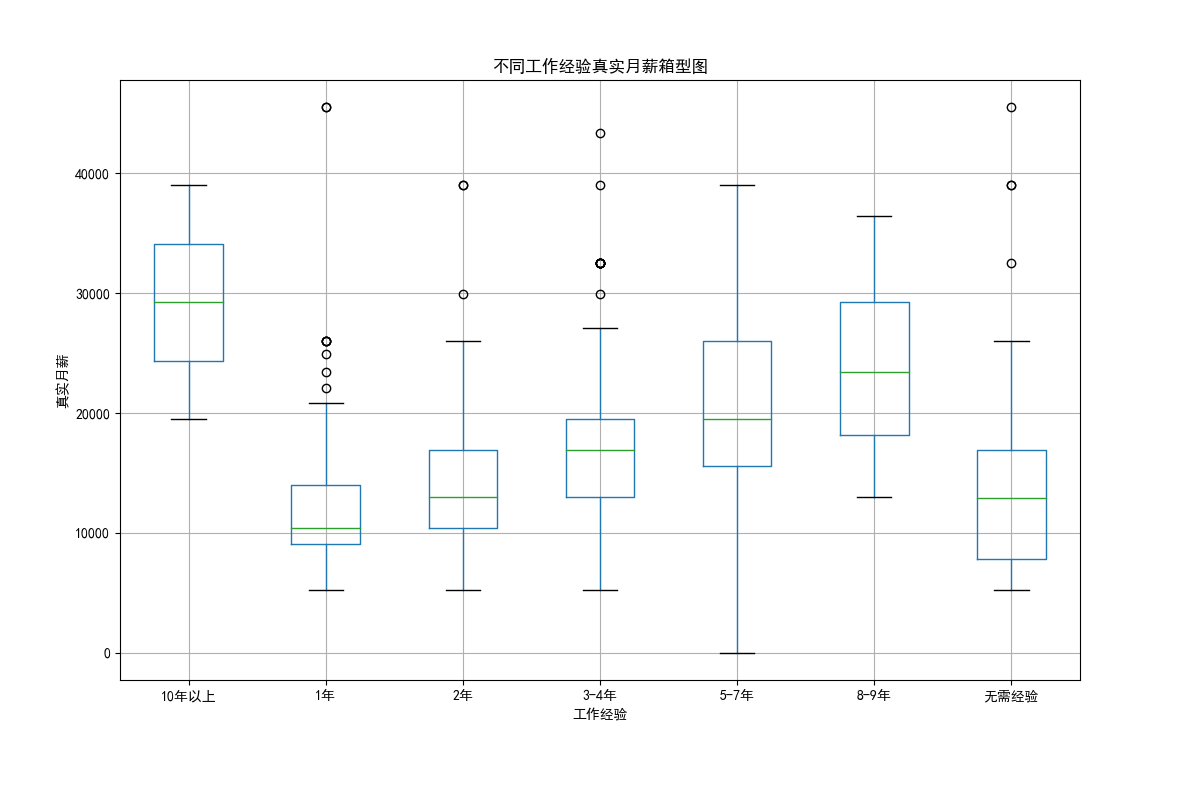
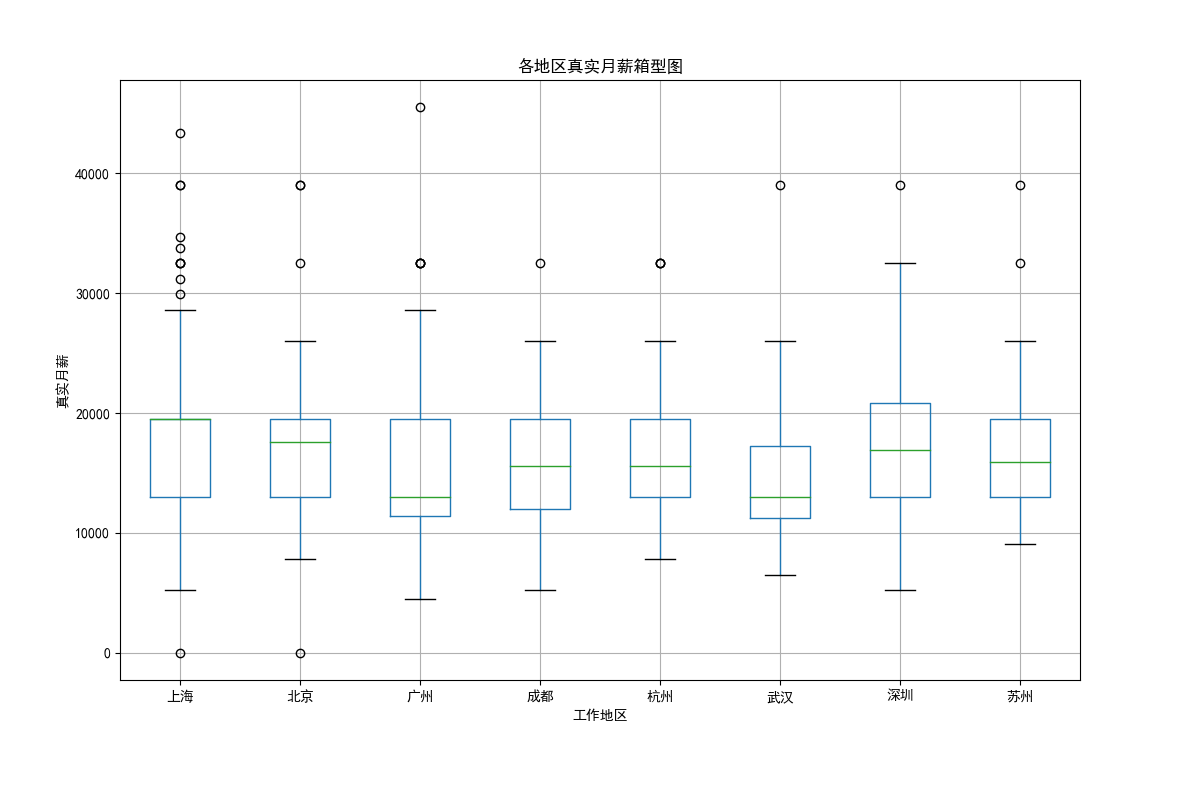
1. 统计了HR在招聘软件上发布招聘信息时最喜欢提到的关键词并生成了一个词云，在词云中可以看到，除了“python”这个关键词外HR提到最多的三个词分别是“五险一金”，“绩效奖金”，“年终奖金”由此可以分析出如果公司真的可以提供这些福利那说明python相关岗位的福利待遇还是不错的。

**PS.这里附上相关代码**

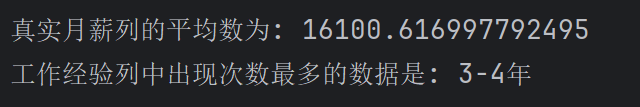
1. 本项目的重点功能和存在的最大意义就是到底python相关岗的薪资到底是多少，如果没有这部分分析那本项目将没有实际意义，在使用爬虫工具爬取到数据后，得到的数据非常混乱，尤其是最关键的“薪资”数据，各个HR在发布信息时薪资的格式都不统一，格式包括但不限于如：“1-2万·13薪”，“7千-1.3万”，“7000-1.2万·15薪”，“1200元/天”等等等等...。但也不是完全没有规律，比如拿现实中最常见的例子举例，如果一个岗位的月薪是“5000-8000”，那大概率就是5000/月，而且这些薪资共有的一个规律就是开头的第一个“数字”一定是底薪，虽然单位不同但是可以用正则表达式和选择语句解决，为了数据更贴近真实，选择将底薪乘以1.3（包括有些公司真的会给的工资接近最大值并且包含如14薪这类情况，而此项目计算的是月薪就需要考虑这些问题），在经过一系列处理后数据终于是能够进行分析了。

**PS.这里附上处理前后的对比图**



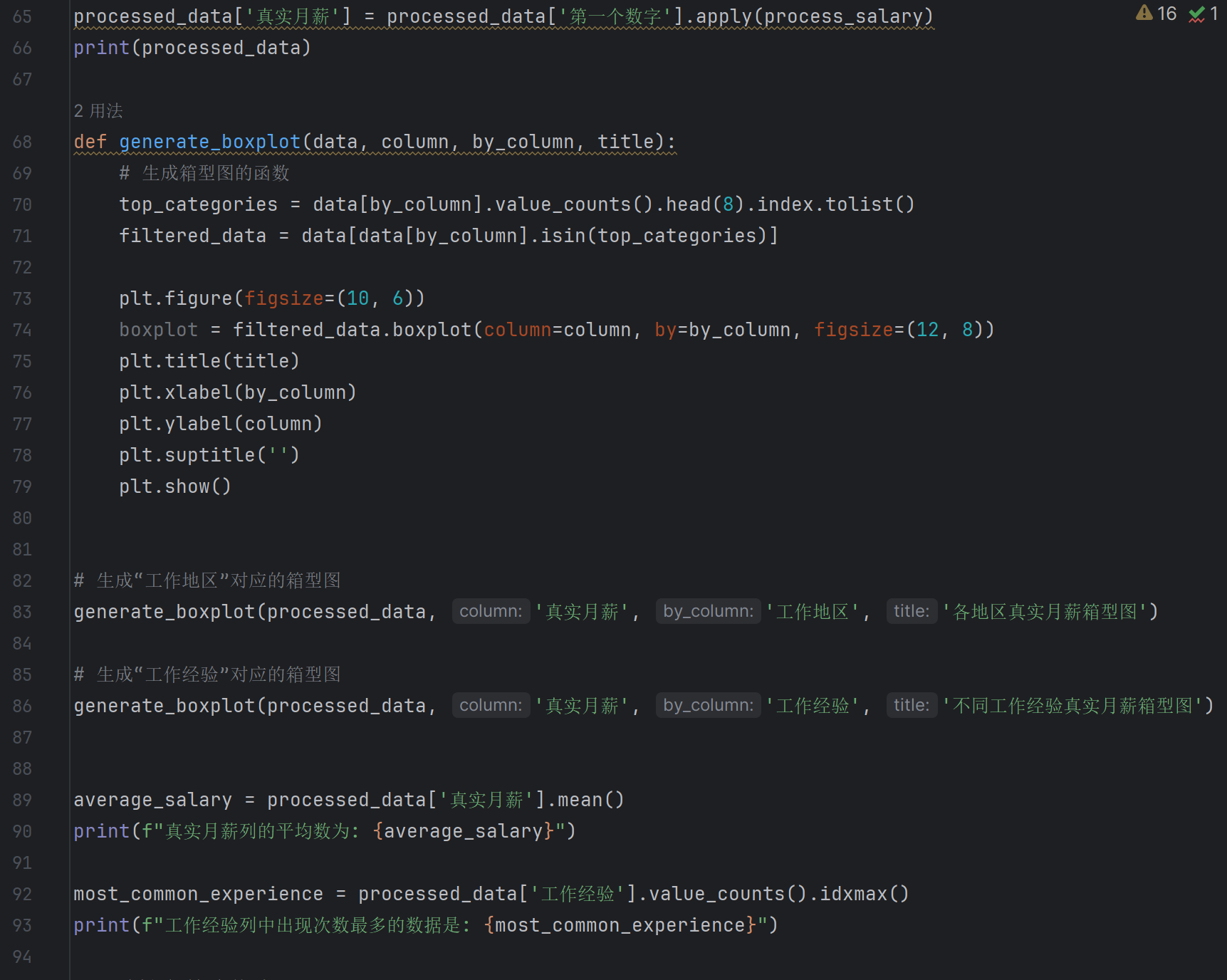
此项目主要将薪资与两个关键条件进行分析，分别是“地区”和“工作经验”（选区的数据都是各自数据列中的前八名），由箱型图可知薪资在每个“热门地区”的大体薪资都是差不多的，基本都在1-2万每月，其中上海，北京在2万附近，广州，武汉在1万附近，其他地区都在1.5万附近。不同工作经验的薪资总体是工作年限越多，工资越高这个趋势（有些比较离谱的，违背常理的数据是因为在数据清洗的过程中难免会有纰漏），从整体上看python相关岗位的薪资待遇还是相对高的，从数据上看经验丰富，并在经济发达的地区工作的工资大部分能够达到3万/月，即使是刚刚毕业只有实习经历，或者没有经验的应届生的工资也有大几千到1万元。

最后还分析了一下python相关岗的平均月薪和公司最喜欢招聘的工作经验的求职者的数据，可以看出来果然大部分公司还是喜欢既年轻，又有足够工作经验的求职者。



**PS.这里附上关键代码**





### 创新功能及新技术实现

这个项目更注重实用性和清洗及分析数据的方法，在做项目的过程中用到了一个很好用的在pycharm中查看CSV文件的插件“ExcelEditor”，还有就是“八爪鱼”这个可视化爬虫软件，它的优势就是可以让完全没有爬虫，甚至是编程基础的用户可以轻松的爬到数据，并且不用担心任何风险。还用到了课上没练过的其他一些文件操作的相关方法和可视化图形，又复习了一些基础的python语法和处理各种奇奇怪怪报错的方法。