ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0026

# Research on Recruitment Information Data Analysis Based on Python Crawlers

Guozheng Hou[1, *, #], Li Tian[2, #], Jing Hou[3], Wenjing Liu[1]

[1]Chongqing Health College, Chongqing, 400000, China

[2]Laifeng Secondary Vocational School, Hubei, Enshi, 445000, China

[3]Guangdong University of Science and Technology, Guangdong, Dongguan, 523000, China

[*]Corresponding author: 15971713575@163.com

[#]These authors contributed equally.

## Abstract

**The current teacher recruitment market presents an imbalance between supply and demand, and this contradiction between supply and demand not only affects the quality of school education but also puts pressure on the career development of teachers and the recruitment market. This study uses Python web crawler technology to crawl physical education teacher recruitment information and extract data from it. After visualization, analyze the data on salary and benefit levels, city demand, education, and ability requirements. The study found that in our country for physical education teacher's education is not high, undergraduate and specialist opportunity positions accounted for 44% and 45% of the total number of positions, respectively, and master's degree positions accounted for only 5%. In terms of salary, there is not much difference between a college degree and a bachelor's degree. In contrast, the average salary of a master's degree is 41% higher than the average salary of a bachelor's degree. Its lower limit is only 24.90% more, but the upper limit is 1.5 times higher than the upper limit of undergraduate salary. First-tier cities pay more attention to physical education and offer superior salaries, especially Beijing, Shanghai, and Guangzhou. In addition, coastal cities usually have more job opportunities than cities in the central region.**

## Keywords

**Python; Web Crawler; Internet; Data Analysis; Visualization.**

## 1. Introduction

With the rapid development of Internet technology, the way of information access is experiencing unprecedented changes. Massive online data provides rich information resources for various industries, but how to extract valuable data from it has become an urgent problem[1]. In the field of human resources, the recruitment market serves as an important bridge connecting enterprises and job seekers, and its dynamic changes have an important impact on social employment and economic development. However, the traditional way of collecting recruitment information often faces the problems of slow data updating and limited coverage, making it difficult to fully reflect the market demand[2]. In addition, for job seekers, it is inefficient to browse job websites and manually collect recruitment-related information, easy to miss key information, easy to be interfered with by invalid or erroneous information, and cannot rationally analyze the average quantitative level of salary and benefit level, city demand, education and ability requirements of the positions from the perspective of the industry. Therefore, the use of web crawler technology to quickly obtain and analyze recruitment information has become an important means of research. Through the web crawler

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0026

technology then can effectively avoid the above defects, fast retrieval of massive information, accurate mining, optimize people's network experience, and save time and energy[3].

As data scientist Hilary Mason said, "Data is the new oil.", Python crawler technology is just the rig to mine this valuable resource. A crawler program written in Python can automatically capture a large amount of recruitment data from the Internet[4]. In this paper, Python crawler technology to take the recruitment information of physical education teachers in "Worry-Free.com" as the data source, and use Python crawler to obtain the data, and then process the data, visualize and analyze it, and analyze the salary, benefit level, city demand, education level, and so on. Then processed and visualized the data, analyzed the salary, welfare level, city demand, education and ability requirements, etc., and tried to reflect the real situation of the current physical education teacher recruitment market.

## 2. Data Acquisition

In this paper, the crawler script calls the Request module in Python to issue a request to the corresponding URL, get the Response response, and the Response contains all the information of the whole page, through the Re regular expression to parse and crawl the study of all the data needed to contain the recruitment information, and then use the built-in Python CSV function to store the final data in an Excel table for visual analysis and display. Excel table for visual analysis and display. The overall process is shown in Figure 1.


**Figure 1.** Crawler operation flow

Open "Worry-Free", press F12 to enter the developer mode to observe, find that the Request Method is getting mode, Status Code is 200, and get User-Agent, Accept-Language, Connection, and other important parameters, at the same time to prevent being blocked as a robot, this paper will store several groups of headers in a list, each run by randomly selecting a group, greatly improving the crawler script. At the same time, to prevent being blocked as a robot, this paper stores multiple groups of headers in a list and randomly selects a group by random each time it runs, which greatly improves the security and stability of the crawler script.

On the page of job posting information, according to the desired data type, get the text response of the job posting information through response.text, use regular expression find all to pinpoint the string containing the desired information, and use JSON.loads to decode the acquired string to return the Python field. Next, create an empty list and iterate through it to get information such as company name, work location, educational requirements, salary, and benefits. Then operate with Python's built-in CSV function to write the data to a CSV file to realize the data-saving task.

## 3. Data Preprocessing

Eleven fields were crawled this time, as shown in Table 1.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0026

**Table 1.** Crawling data fields

| Parameters | data field |
|---|---|
| 'Job Title' | title |
| 'Basic information' | jjj |
| 'Company Name' | company_name |
| 'Location' | workarea_text |
| 'Company Type' | companyind_text |
| 'Company Size' | companysize_text |
| 'Company Nature' | companytype_text |
| 'Benefits' | jobwelf |
| 'Salary' | providesalary_text |
| 'Posting time' | updatedate |
| 'Job details page' | job_href |

### 3.1. Data Cleaning

Data cleaning is preparatory work that must be completed before visualization and analysis, i.e., removing duplicates, errors, and blank information. Such data will affect the subsequent analysis, so the steps of data cleaning are very important. Data cleaning mainly includes the following steps: (1) delete error data; (2) deal with duplicate data; (3) deal with null values; (4) detect outliers and deal with abnormal data. For the null values appearing in the salary, if all of them are deleted it will reduce the reliability of data analysis, find out the position where the missing data are located, traverse the salary data of the position, and get the multitude of values to fill in the null values. Finally, delete the rows where the other null values are generated during data collection[5].

### 3.2. Normalization process

Salary, data exists in the form of both range values, but also appears a variety of different units of measurement. To facilitate analysis, the existence of the range form of data for the maximum value, minimum value, and average value of the processing, at the same time, the "thousand dollars/month" as a unified form of monthly salary, adding the highest salary, minimum salary, and average salary.

## 4. Data Visualization Analysis

### 4.1. Analysis of Job Posting Time

This paper crawls the data from late February to late April 2022, a total of 60 days (Wuyou Network). As shown in Figure 2, the amount of information released from February to mid-April has been stable, and with the advent of the “graduation season”, the amount of information released in late April suddenly increased dramatically, and the number of job information released on April 22 was as high as 89, ranking first. Since then, the number of job postings has increased to a much higher frequency than usual (8 postings/day). From the analysis of the data, it can be seen that there is a positive relationship between the market job postings and the demand, which to a certain extent protects the needs of graduates looking for jobs. However, the data also reflects the lack of job postings during the non-graduation season, and the piling up of job postings at the same time, which increases the pressure of job hunting, is not conducive to the reasonable distribution of labor, and wastes human resources.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0026

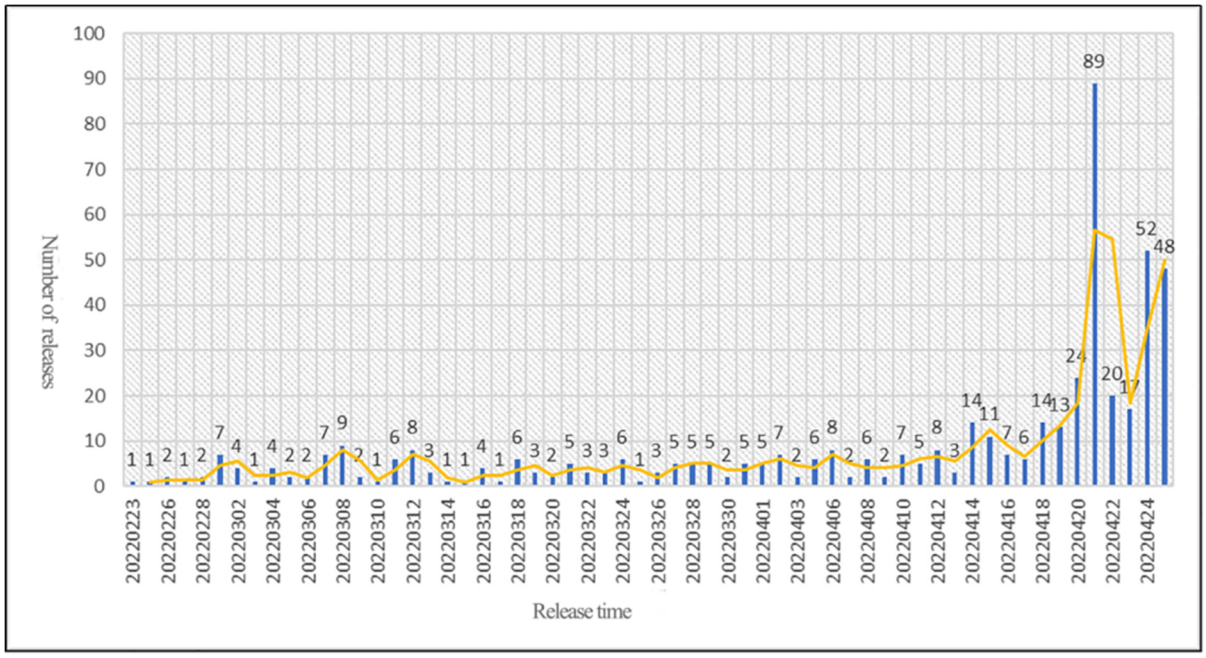


**Figure 2.** Recruitment information release time

## 4.2. Analysis of Academic Qualifications

In the sports industry, there is the selection and assessment of sports skills in the recruitment information, which somewhat weakens the influence of academic qualifications on job seeking. As can be seen from Figure 3, China's educational requirements for physical education teachers are not very high. The opportunity positions for undergraduates and specialists accounted for 44% and 45% of the total number of positions respectively, with a difference of only 1%. The positions requiring a master's degree are only 5%, and the high degree of education does not occupy a great advantage in the sports market, this data reflects to a certain extent that the market demand for physical education teachers' education is concentrated in bachelor's degree and specialization.

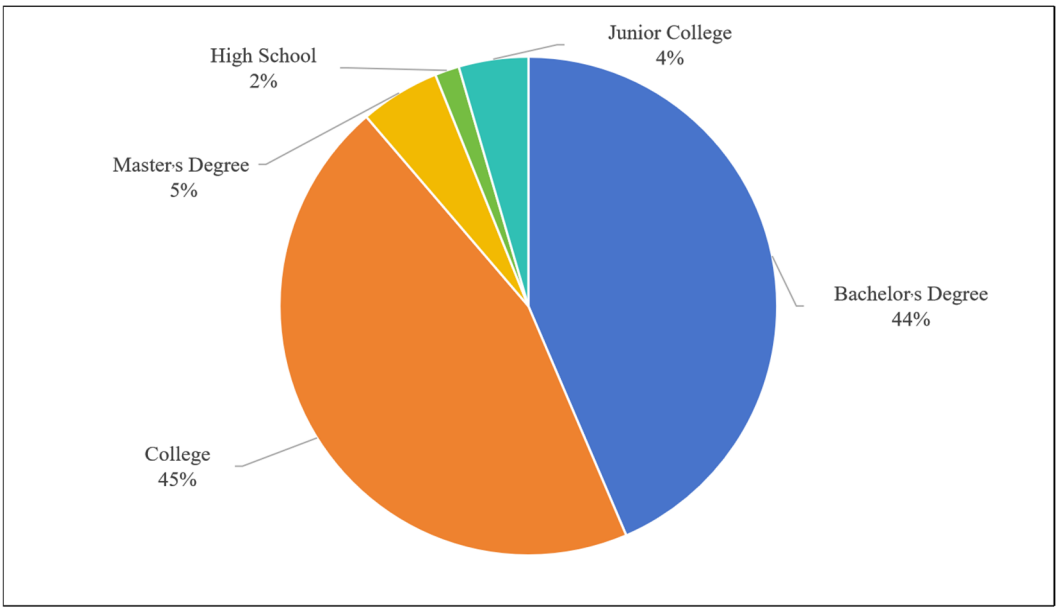


**Figure 3.** Educational Requirements for Physical Education Teachers

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0026

## 4.3. Analysis of salary situation

The distribution of salaries for different academic degrees is shown in Figure 4. There is not much difference between the salary package of a college degree and a bachelor's degree, while the average salary of a master's degree is 41% higher than the average salary of a bachelor's degree, and its lower limit is only 24.90% more, but the upper limit is 1.5 times of the upper limit of bachelor's degree salary. The supply of higher-education jobs is relatively low, but they are higher in salary than undergraduate and specialized degrees. Higher qualifications often mean more in-depth professional knowledge and teaching ability, as well as the possibility of taking up more teaching and research tasks. In addition, some schools or educational institutions offer additional pay incentives for teachers with advanced degrees. This may account for the fact that master's degree salaries are higher than those of bachelor's and specialist degrees.

In addition, as shown in Figure 5. A side-by-side comparison of the average salary of physical education teachers in eight popular cities, including Beijing, Shanghai, Guangzhou, and Shenzhen, with the average salary in each city reveals that the first-tier cities attach more importance to physical education and have superior salaries, especially the average salary of physical education teachers in three cities, namely Beijing, Shanghai and Guangzhou, which is comparable to the average salary in the local area. Salary levels are affected by region, type of school, teachers' experience, teaching performance, and the specific salary structure of the education system in which they work. Therefore, while education is a factor affecting salary, it is not the only determining factor.

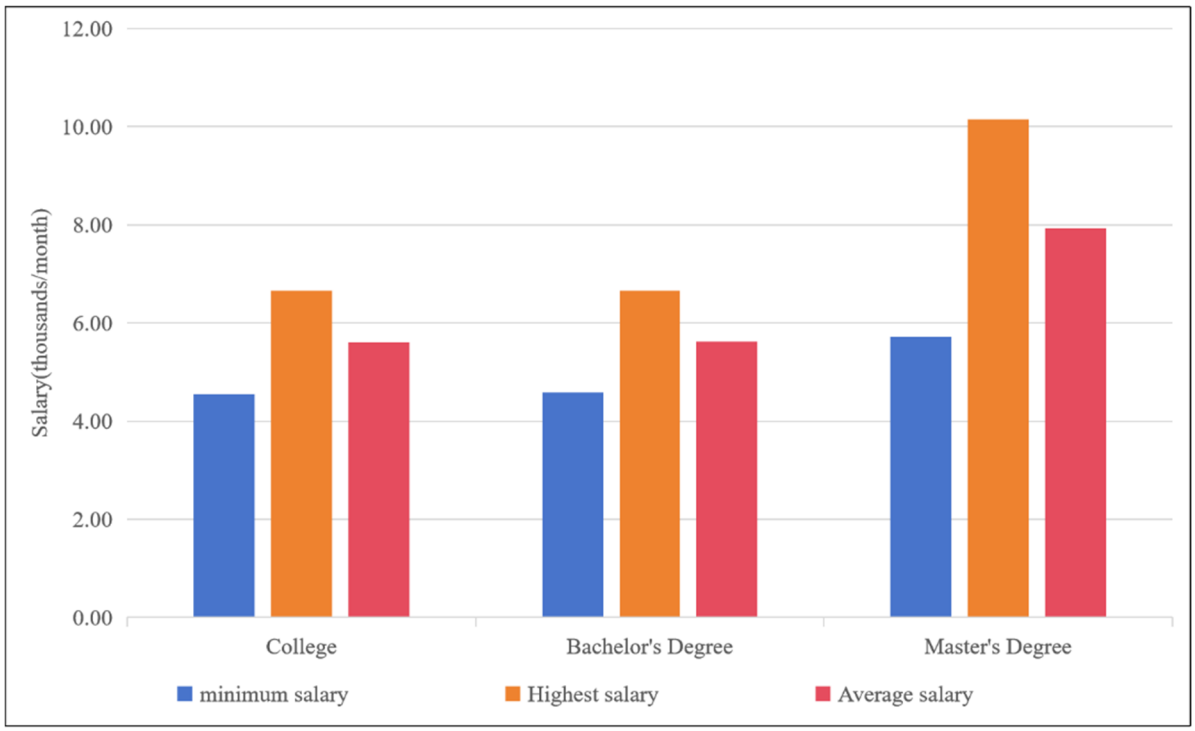


**Figure 4**. Salary analysis by education level

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0026

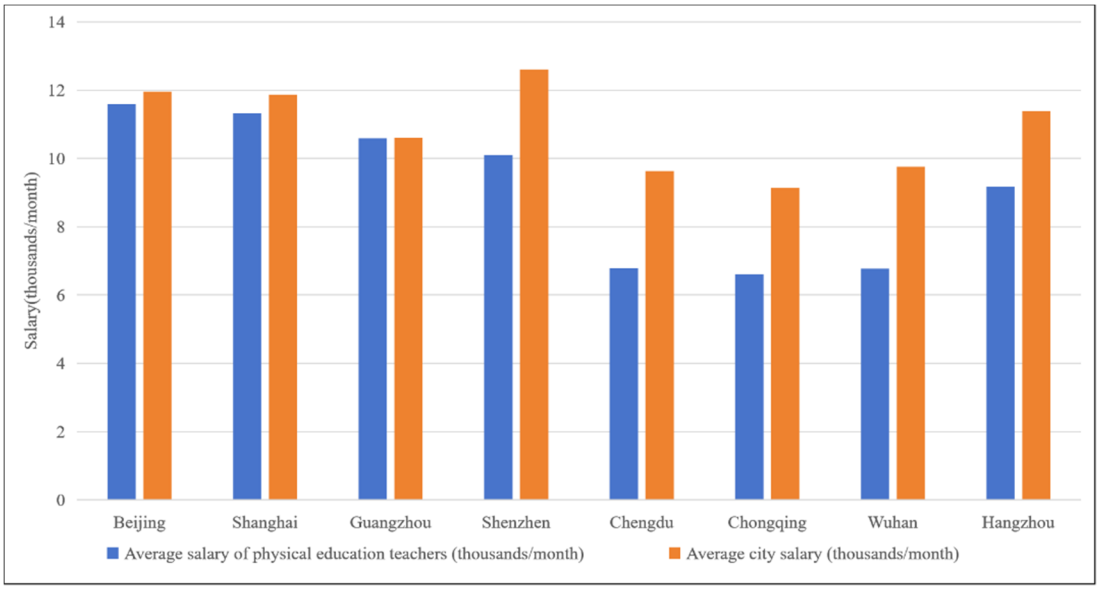


**Figure 5.** Comparison of Average Salary of Physical Education Teachers in Different Cities

## 4.4. Analysis of Urban Demand

With the increasing emphasis on physical education, the demand for physical education teachers is growing in various cities, and the recruitment of physical education teachers in major cities is shown in Figure 6. The data shows that first-tier cities such as North, South, and South China have a high demand for physical education teachers and have more employment opportunities. This may be related to the fact that economically developed regions have more educational resources and can provide more employment channels. From the perspective of geographical distribution, it can be seen that there are usually more job opportunities in coastal cities than in cities in the central region, on the contrary, Wuhan has 42 job openings, which is second only to the first-tier cities. This may be due to the influence of the regional culture of Wuhan on the demand for physical education teachers, as Wuhan is relatively part of a region that emphasizes sports, so the demand for physical education teachers increases, indirectly providing more job offers. In addition, the climatic conditions and the degree of sports facilities in the region also affect the working environment and teaching effectiveness of physical education teachers, which in turn affects the recruitment situation.

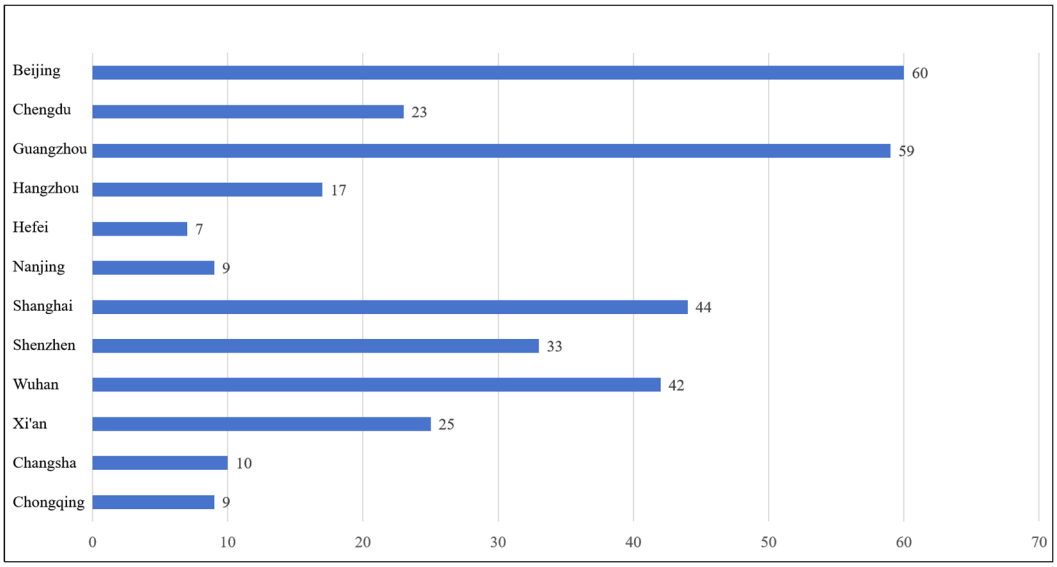


**Figure 6.** Recruitment by major cities

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0026

## 5. Conclusion and Recommendations

In this study, using Python crawler technology, it was found that the demand for physical education teachers varies significantly across regions. The demand for physical education teachers in first-tier cities and coastal developed regions is significantly higher than that in inland and less developed regions. This is closely related to the level of economic development and the degree of importance attached to physical education; in addition, the salary level of physical education teachers has a positive correlation with the level of regional economic development, and this positive correlation is not limited by academic qualifications; this study also found that in the job postings of physical education teachers, high academic qualifications can be given a higher salary, but there is no advantage in the supply of job postings, and the supply of job postings is even lower relative to undergraduate and professional qualifications.

In response to the results of this study, it is recommended that education departments and schools adopt differentiated geographical recruitment plans in their recruitment strategies to meet the demand for physical education teachers in different regions. Meanwhile, the data analysis also revealed a positive correlation between PE teachers' salary and recruitment demand, i.e., districts with higher salary levels tend to attract more PE teachers. Therefore, it is recommended that consideration should be given to increasing the salary packages of physical education teachers when formulating recruitment budgets to attract and retain outstanding talents.

## References

[1] Chien C-F, Chen L-F: Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry[J]. Expert Systems with Applications 2008, 34:280-90.

[2] Ge C, Shi H, Jiang J, Xu X: Investigating the Demand for Blockchain Talents in the Recruitment Market: Evidence from Topic Modeling Analysis on Job Postings[J]. Information & Management 2022, 59:103513.

[3] Karakatsanis I, AlKhader W, MacCrory F, Alibasic A, Omar MA, Aung Z, Woon WL: Data mining approach to monitoring the requirements of the job market: A case study[J]. Information Systems 2017, 65:1-6.

[4] Goldfarb A, Taska B, Teodoridis F: Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings[J]. Research Policy 2023, 52:104653.

[5] Yang S, Abas A: Data Science Talents Mining from Online Recruitment Market in China Based on Data Mining Technique[J]. Journal of ICT In Education 2021, 8:118-25.