Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 19, No. 1, 2025

Research on supply and demand characteristics of big data science industry based on multi-dimensional analysis

Gaorui Zhang [a], Huiqin Sun [b], Xiaohong Wang*

School of Information Engineering, Wuhan Business University, Wuhan 430056, China
* **Corresponding author Email:** 6934296@qq.com, [a]2486846653@qq.com, [b]3498065134@qq.com

Abstract: The present study aims to investigate the employment prospects of Data Science and Big Data Technology majors. To this end, a web crawler system has been constructed using Python technology. The software extracts data on "Data Analyst" positions from a recruitment website, performs operations such as data structuring, duplicate removal, salary outlier handling, and standardization of educational requirements. The integration of descriptive statistics and visualization techniques facilitates the establishment of a comprehensive database, encompassing variables such as salary, geographical location, educational attainment, and skillset. Empirical analysis reveals significant regional disparities in salary for data analyst positions, with Beijing, Shanghai, and Shenzhen averaging 25-35K RMB monthly—30%-50% higher than other cities. A positive correlation is evident between educational attainment and salary, with doctoral degree holders earning approximately 1.8-2.2 times the average monthly salary of bachelors degree holders. It is evident that Python, SQL and Tableau are the skills most frequently mentioned among the skill requirements, with percentages of 95%, 92% and 82%, respectively. The findings of this study provide data-driven insights with regard to the development of academic programs and the planning of careers.

Keywords: Crawlers; Data analysis; Data visualization; Data analysts.

1. Introduction

In the contemporary era of big data and the proliferation of the internet, online recruitment has become an integral component of university hiring practices. This innovative approach to recruitment has had a significant global impact, providing universities with a pathway to further reform their hiring activities while also granting them a competitive edge in attracting top talent. This initiative establishes a fundamental framework for the future stability and rapid development of universities, while also creating avenues for job seekers to leverage fragmented time for precise job matching [1].

Data science and big data technology, as emerging interdisciplinary fields, find applications across various positions [2]. It is imperative to note that research focusing on talent cultivation and the supply-demand dynamics in the job market is of particular significance. However, even within the same profession, there is significant variation in salary levels across different cities and educational backgrounds. The objective of this paper is to explore the patterns of talent supply and demand for data analyst positions, and the underlying logic supporting industry development. The objective of the initiative is to enhance the alignment between the cultivation of professional talent and the growth of the industry. By analysing recruitment data across various dimensions, including city distribution, educational requirements, skill demands, and work experience, the study explores the intrinsic connection between job characteristics and industry evolution. This research provides valuable direction for job seekers in this field.

2. Data sources and processing

2.1. 1.1Data sources

The data of this study comes from a recruitment website, in which the position of "data analyst" is selected as the research object, and Python-based crawler technology is used to obtain the relevant recruitment data such as city, job salary, educational requirements, etc., and then the acquired data are subjected to the data structured conversion, duplicate data removal, salary and outlier processing, educational requirements standardization and other operations. outliers processing, standardization of educational requirements and other operations.

2.1.1. Crawler parameter settings

After designing the technical route, according to the target data, the city in the parameter is set as "National", the educational requirement is set as "Unlimited", the position is set as "Data Analyst", the working experience is set as "Unlimited", and the position is set as "Full-time". ", work experience is set to "unlimited", and the position is set to "full-time". Crawler parameters are set as shown in Table 1.

Table 1. Crawler parameter settings

parameters	numerical value
city	nationwide
position	full-time job
education attainment	not limited to
office	Data Analyst
working experience	not limited to

2.1.2. Data Acquisition

The HTTP protocol-based web crawler technology is used to simulate browser behavior by constructing standardized request headers: the request header contains randomly generated User-Agent strings to simulate different terminal

devices, and supplemented with fields such as Accept, Accept-Language, and Referer to ensure the compliance of the request format[3]. With "data analyst" as the keyword, GET requests are sent to the target recruitment platform according to the paging parameter (40 records per page), and the random access interval of 2-4 seconds is set to circumvent the server anti-climbing mechanism by controlling the request frequency. The response content is parsed to extract 12 items of structured data, such as job title, salary range, working city, educational requirements, working experience, skill label, company information and release time, and the cumulative sample size obtained meets the minimum requirements for statistical analysis.

2.2. Data pre-processing

Data preprocessing is a key link to guarantee the accuracy of the analysis results, the specific steps are as follows (e.g., Figure 1):

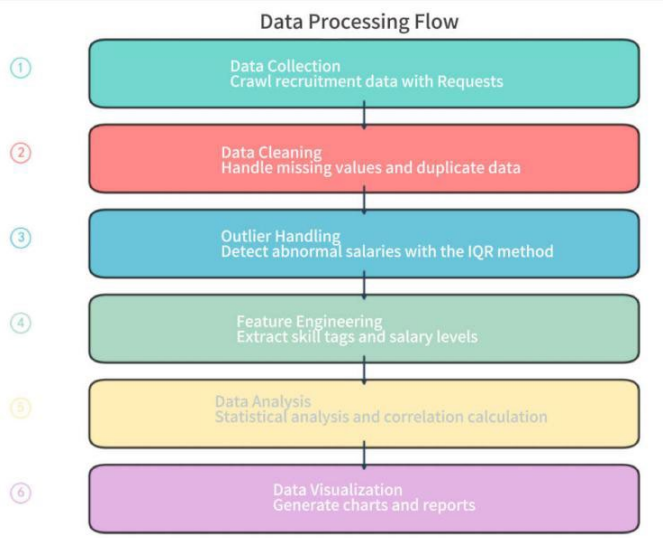

Fig. 1 Flow related to data preprocessing

(1) Data Preparation Phase: Initialization of the data environment is the first step. For instance, in the absence of a DataFrame, a sample dataset is automatically generated. This step ensures subsequent analysis processes have a complete and reliable data foundation, thereby preventing interruptions due to missing data and providing essential input assurance for the entire data processing workflow.

(2) Data deduplication processing [4]: Deduplicate job information based on the job_id field to effectively identify and remove completely duplicate data records. This process significantly reduces data redundancy. By comparing data volume changes before and after deduplication, it visually demonstrates the data cleansing effect, providing a cleaner dataset for subsequent analysis.

(3) Outlier Detection and Handling[5]: By establishing a reasonable salary range standard (3,000–80,000 RMB), the system filters and removes outlier salary records exceeding this range. This step effectively eliminates interference from unrealistic salary data, ensuring the accuracy and credibility of analysis results while maintaining overall data quality.

(4) Data standardization: Data standardization is achieved through the establishment of the academic degree mapping table, which unifies the expressions of academic degrees such as "college" and "bachelors degree" into standard classifications. The missing values are filled with "unlimited" consistency, so that all the academic qualification data formats remain uniform, and the final output is a clean and regular dataset.

(5) Data quality assessment: Conduct a comprehensive quality assessment of the dataset, including completeness, accuracy and consistency checks. Data quality is verified through statistical indicators and visualization methods to ensure that the processed data meet the analysis requirements.

(6) Feature engineering construction: Construct analytical features based on the cleaned data, including standardized processing of numerical features and coding conversion of category features.

(7) Data archiving and storage: normalize and store the final processed data set, establish version management and metadata records.

3. Data visualization and analysis

Visual analysis is an important method of big data analysis [6]. Big data visual analysis aims to utilize the automated analysis capability of computers while fully tapping the cognitive ability advantage of people for visual information, organically integrating the respective strengths of people and machines, and assisting people to learn more intuitively and efficiently about the information, knowledge and wisdom behind the big data with the help of human-computer interactive analysis methods and interactive technologies.[7]

3.1. Analysis of the relationship between city and average salary

As shown in Figure 2, data analyst positions exhibit an imbalance between supply and salary across cities. In terms of job distribution, Nanjing ranks first with a 14.0% share, followed by Beijing and Hangzhou at 11.5% each, Suzhou at 12.0%, Xian at 10.5%, Chengdu at 10.0%, Guangzhou and Wuhan at 9.0% each, while Shanghai and Shenzhen account for only 6.5% and 6.0% respectively. However, in terms of salary, Shenzhen and Shanghai lead with the highest average wages, followed by new first-tier cities like Beijing and Hangzhou. Despite accounting for 10.5% of job distribution, Xians average salary is only ¥13,000.00. Overall, economic hubs like Shenzhen and Shanghai exhibit low job distribution shares yet strong salary competitiveness, while cities such as Nanjing and Xian offer abundant job opportunities but weaker salary levels. This reflects how regional industrial structures and varying talent demand tiers significantly influence job compensation.

As demonstrated in Figure 3, the minimum salary for data analyst positions across cities exhibits both hierarchical differentiation and heterogeneous dispersion. With regard to the median minimum salary, Shenzhen is the top-ranked city, at approximately ¥19,800, followed by Shanghai. Hangzhou and Guangzhou occupy the middle tier, while central-western cities such as Chengdu, Wuhan, and Xian demonstrate lower levels. Guangzhou occupies the middle tier, while central and western cities such as Chengdu, Wuhan, and Xian exhibit lower levels. This finding indicates a descending salary hierarchy pattern of "first-tier cities → new first-tier cities → central and western cities," which is consistent with the gradient differences in urban economic capacity and digital industry concentration. With regard to dispersion, Shenzhen and Shanghai exhibit wide box spans, extensive whisker extensions, and numerous outliers, reflecting significant internal variation in minimum salaries. This variation is characterised by both high baseline "floor salaries"

supporting complex scenarios and differentiated entry-level wages driven by diverse industrial demands. In contrast, Hangzhou, Guangzhou, and central-western cities exhibit progressively narrower boxes, shorter whiskers, and fewer outliers. It is evident that cities such as Chengdu, Wuhan and Xian exhibit compact boxes, suggesting a more concentrated minimum wage distribution with reduced volatility and differentiation.

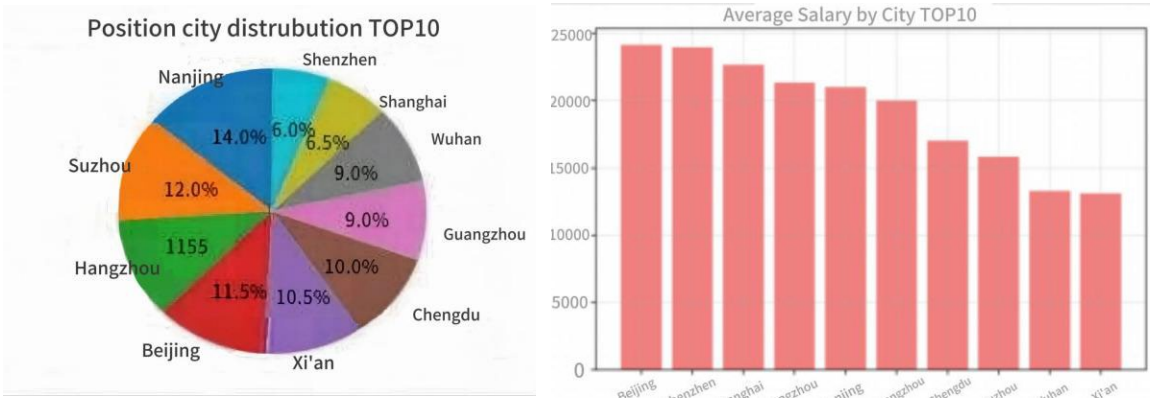

Fig. 2 Top 10 Distribution of Post Cities

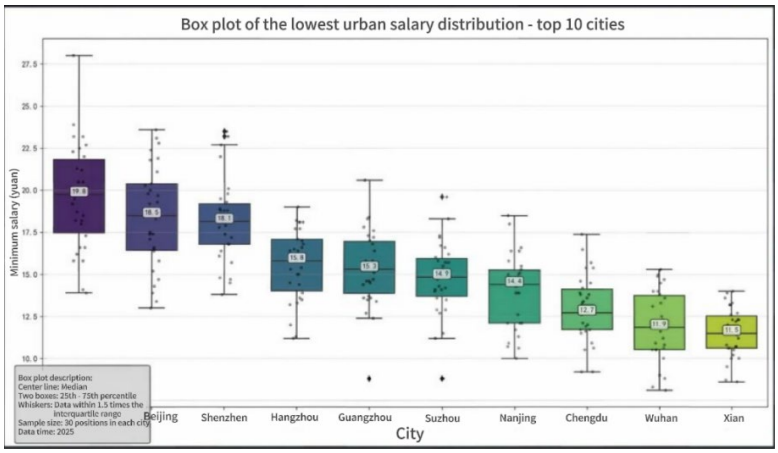

Fig.3 Top 10 Average Salary Box Plots by City

3.2. Analysis of the relationship between academic requirements and average salary

As demonstrated in Figure 4, there is a substantial positive correlation between the educational requirements for data analyst positions and average salaries, with clear and pronounced differences in compensation levels across the various educational tiers. The mean salary for individuals in possession of a doctoral degree approaches ¥30,000, thus placing them in an absolute leading position among all educational levels —nearly 2.5 times that of individuals in possession of an associate degree. It is evident that those in possession of a masters degree accrue an average income of approximately ¥20,000, a figure that clearly surpasses that of individuals with either a bachelors or associate degree. Bachelors degree holders have an average income of approximately ¥15,000, while those with an associate degree earn around ¥12,000. This finding suggests that higher levels of educational attainment may confer data analysts with greater salary bargaining power in the job market. This is indicative of the industrys high recognition of the knowledge, skills, research and innovation capabilities of highly educated professionals. In addition, the findings provide empirical evidence in support of the human capital theory. Specifically, the results demonstrate that educational qualifications serve as a pivotal conduit for human capital accumulation, thereby exerting a direct influence on salary returns. This underscores the positions rigorous prerequisites for the depth of knowledge reserves and professional development potential of its talent pool.

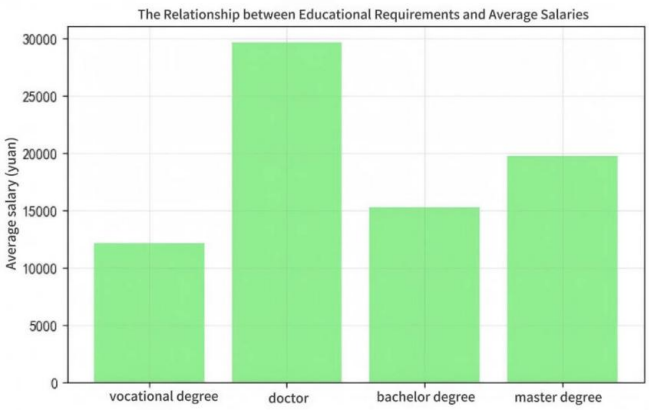

Fig. 4 Histogram of the relationship between educational requirements and average salary

3.3. Skill demand relationship analysis

As illustrated in Table 2, there is a clear delineation in the market value of data analyst skills. Python, Power BI and Tableau represent core competencies characterised by high pay and high demand, with both salary levels and demand intensity predominantly rated as high. The market competitiveness of these professionals has been determined to be at the levels of "extremely high" or "high," reflecting the industrys strong demand for skills that combine technical depth with broad applicability, resulting in significant skill

premiums. SQL and Excel, as foundational tools for data work, are in high demand. However, due to their widespread adoption and relatively low technical barriers, their salary levels tend towards "low" or "medium," with market competitiveness rated as "moderate." These competencies are considered to be more essential for the role, and as such, they offer limited potential for salary growth. R, SAS, and SPSS demonstrate relatively balanced performance. The majority of these positions are characterised by median or low salaries, and moderate levels of market competitiveness. These skills are in demand for specialised scenarios, such as statistical analysis and industry-specific modelling. However, their application breadth or technical uniqueness falls short of the first two categories, resulting in overall mid-range market value.

Table 2. Top 15 Data Analyst Skill Needs

skill name	Average salary ($)	Frequency of demand (%)	pay scale	Level of demand	Market Competitiveness
Python	27,932	95%	high	high	extremely high
Power BI	29,881	88%	high	high	extremely high
SQL	21,622	92%	low	high	moderate
R language	25,941	78%	moderate	moderate	moderate
Excel	23,374	85%	moderate	moderate	moderate
SPSS	22,153	65%	low	low	moderate
Tableau	27,977	82%	high	high	high
SAS	26,900	75%	moderate	low	moderate

As shown in Figure 5, the weight difference of skill demand is presented visually through font size: Tableau, Hadoop and other fonts are the most significant, forming a closed loop with the high demand for Tableau (33.5%) and big data tools (e.g., Spark 33.0%); the words "data visualization" "Machine Learning" and other fonts also occupy the same proportion as the corresponding part of Table 2, more intuitively reflecting the industrys high demand for data processing, data analysis and modeling.

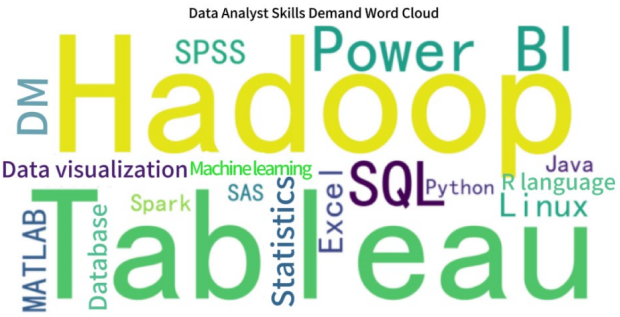

Fig. 5 Word cloud of data analyst skill requirements

3.4. Analysis of the relationship between work experience and average salary

As demonstrated in Table 3, an analysis of the data indicates a substantial correlation between work experience and the average salary for data analyst positions. Those with five to ten years of experience earn an average salary of ¥29,046, which is approximately 2.8 times that of recent graduates. This reflects the industrys premium for seasoned talent. It is evident that there is a direct correlation between the duration of experience in a role and the subsequent increase in salary. Individuals with a proven track record of three to five years experience receive a 25% salary increase compared to those with one to three years experience. This data serves as a reliable indicator of the value accumulation of technical proficiency and project experience. It has been demonstrated that positions which do not require prior experience tend to offer slightly higher salaries than those which require between one and three years of experience. This discrepancy can potentially be attributed to the more complex skill sets demanded by the latter roles. Despite the modest levels of initial remuneration, an analysis of the skill requirements (with a high prevalence of Python, Spark, etc.) indicates that the industry places a premium on practical capabilities. The prevailing tendency is evidenced by the finding that accumulated experience is associated with salary advantages, which are indicative of enhanced professional depth and industry insight.

Table 3. Relationship between work experience and average salary

working experience	Average salary (dollars)
5-10 years	29046.51
3–5 years	21151.16
1–3 years	16885.71
Unlimited experience	15602.04
student of the current year	10333.33

4. Conclusion

The present study employs a multidimensional analysis of data analyst job postings to reveal market patterns regarding urban distribution, educational requirements, skill demands, and work experience. The urban dimension exhibits an imbalance between supply and salary, reflecting regional disparities in digital industry structures. The findings demonstrate a strong correlation between education, experience, and salary, suggesting a positive relationship between these factors and compensation. The industrys predilection for composite competencies, integrating "tool operation + analytical methodologies," is underscored by the stratified nature of skill requirements. From an industry development perspective, the high degree of salary variation observed in major cities such as Shenzhen and Shanghai indicates an evolution in data analyst roles, moving from a more standardised, foundational analysis towards a more specialised approach, characterised by "high-end composite + scenario-specific specialisation." The recurrent demand for

tools such as Python and Spark is congruent with prevailing trends in big data processing and visualisation analysis within the digital economy.

The findings for data science and big data technology programmes provide clear guidance for talent development. This guidance suggests that instruction in core competencies such as Tableau visualisation tools and Python programming should be strengthened, whilst also deepening the theoretical foundations in statistics and machine learning. Collaborations between academia and industry have been shown to enhance students practical skills, thereby bridging the gap between educational offerings and industry demands. This approach is expected to facilitate the attainment of mutual professional growth and industry advancement for data analysts in the context of the digital economys ongoing transformation.

Acknowledgements

This paper was funded and supported by the 2024 School-level Innovation and Entrepreneurship Training Program of Wuhan Business University (No.202411654184).

References

[1] Yang Lu. Problems and suggestions analysis of network recruitment in colleges and universities in the era of big data[J]. China Management Informatization,2021,24(19):121-123. DOI:CNKI:SUN:GLXZ.0.2021-19-054.

[2] Shugui Zhang. Design and implementation of Hadoop-based big data platform for smart job analysis[J]. Information and Computer (Theoretical Edition),2024,36(05):112-114+118. DOI:CNKI:SUN:XXDL.0.2024-05-034.

[3] H. Zhang, B. An, J.F. Zhang. Crawling and analyzing the recruitment data of big data professionals based on Python[J]. Journal of Taiyuan City Vocational and Technical College, 2025, (10):76-78.DOI:10.16227/j.cnki.tycs.2025.0619.

[4] Tan YJ. Research on data de-duplication technology in data backup system[D]. Huazhong University of Science and Technology,2012.

[5] Cheng Kaiming. Review of theories and methods of statistical data preprocessing[J]. Statistics and Information Forum,2007, (06):98-103.DOI:CNKI:SUN:TJLT.0.2007-06-020.

[6] H.P. Ding,R.J. Ji,C.C. Zhao,et al. Recruitment data crawling and visualization analysis based on Python[J]. Computer Programming Skills and Maintenance,2025, (08):103-105. DOI:10.16184/j.cnki.comprg.2025.08.017.

[7] Ren Lei,Du Yi,Ma Shuai,et al. An overview of big data visual analytics[J]. Journal of Software,2014,25(09):1909-1936. DOI:10.13328/j.cnki.jos.004645.

[8] Focusing on Digital Economy and Employment--China's Ninety-Eight Human Capital Forum in Xiamen[J]. China Employment, 2022, (10):7-8.DOI:CNKI:SUN:GGJX.0.2022-10-003.