Introduction

This is the first article in the "Career" series of urban data groups. The theme of this series is the career development of migrant workers. We will start from the existing recruitment data and use scientific data processing methods to present the current situation to readers. The occupational income growth curves of different occupations in different cities are also combined with the recent hot spots of public opinion to estimate the occupational substitution that AI may produce.

As the first tweet in the series, I will start with the data used in this series of research and the grouping of "occupations". I hope that after readers understand these basic information and methods, they will be more specific about the follow-up occupational salary and AI replacement rate. The data list has a more accurate understanding.

Recruitment data of the same age as the Internet in China

If you want to ask, among all kinds of Internet data, what type of data is easy to obtain, has a very large amount of information, and has a relatively long look-back time? My answer is - recruitment data.

Go back in time to more than 20 years ago. At that time, most ordinary families did not have their own computers, let alone the Internet, and what was the most attractive online behavior? Not "surfing the web" or "chatting in a chat room". On September 13, 2000 , an article in "Science and Technology Daily" mentioned:

"With the rapid development of our country's economy , the Internet is gradually accepted by the public. According to survey statistics , among the current Internet users , those who use the Internet for the purpose of job hunting account for half of the Internet population."

On October 9, 2000, an article in "Internet Weekly" "Where to find a job in the future?" The article “Comparison of Recruitment Websites” cites a survey by CNNIC in July 2000. Among the information people obtain online, recruitment information accounts for 26.11%.

互联网周刊

46 · 领先半步 · 调查 2000 年 10 月 9 日 责任编辑 张 晔 Email: zhangye@ciweekly.com

未来职业何处寻?

招聘网站大比较

特约调查团 叶远强 李诗扬

随着互联网的快速发展,,网上找工作越来越受到人们的欢迎。与传统的方式——通过职业介绍所、招聘会、熟人介绍以及媒体刊登的广告等途径相比较,网上找工作具有很多显著的优点,如信息传递快、更新快、时效性强、针对性强以及费用低等方面。CNNIC 在 2000 年 7 月进行的调查显示:用户在网上获取的最主要信息方面,求职招聘信息占有 26.11% 的比例;网上信息不能满足用户需要方面,求职招聘信息占有 19.62% 的比例。同时,用户对求职招聘信息的获取也呈上升趋势,1999 年 7 月的统计数据为 19%,到 2000 年 7 月则上升到 26%,用户选择网上找工作,也推动了越来越多的公司将招聘信息发布到网上。据 CNNIC 在 2000 年 1 月的调查结果显示,公司、单位在网上发布的信息中,招聘信息占 26.82%。

表 1

	招聘网	前程无忧工作网	搜狐人才网	搜狐	中华英才网	网易
相关数据	可供查询的数目: 488 家公司 7814 个职位	有效职位 65656 个 空缺职位 288886 个 本月新增职位 25502 个 当日新增 1039 条	现有职位 74979 个 昨日新增 502 个	今日职位 96519 发布兼职工作 893 个 登记兼职人才 11620 人 查询兼职人才 3662 次(昨日) 查询兼职职位 9009 次(昨日)	空缺职位 235141 个 本周新增全职职位 33976 个 英才总数 102584 本周新增全职求职者 2133 个 经理级以上职位 25789 个 应届高校毕业生 11201 个	无

以外,其他五个网站都在主页上提供了相关数据,但这些数据没有一个统一的标准,不能进行相关比较。同时,这些数据的可信度也不能完全保证。我们在9 月 13 日的 15:00 至 16:00 之间,对各个网站提供的数据进行了一次统计,结果见表 1:

由于各网站的数据更新很快,因此表 1 的数据仅仅是我们在评测过程中所记录的结果。其中,中华英才网和前程无忧工

招聘网站的信息中,行业以及地区之间的差异还很明显。就地区而言,广州的信息比北京和上海就少的多。从这次比较的结果看,这无疑是招聘网站的一个通病。

表 2

职位 / 网站	财务会计	软件工程师	文字编辑
招聘网	4	88	3

其境,首先要打造一艘乘风破浪的快艇,这就是通往成功彼岸的通行证——简历。只有当你按照招聘网站提供的简历制作模版将个人信息填写完整,拥有自己的用

In 1997 , ChinaHR and Zhaopin.com were established; In 1999 , 51job was established. In 2005 , 58.com , a job-seeking website for fresh graduates (yingjiesheng), was established. In the mobile Internet era after 2010 , various recruitment websites and apps have sprung up like mushrooms after a spring rain. Liepin, Boss Zhipin, Lagou... expand the dimension of recruitment data again.

From the high-salary offers coveted by elite school students to domestic service and short-term odd jobs for blue-collar workers, the recruitment data has accumulated not only the story of generations of workers, but also the epitome of the Chinese economy over the past two decades.

Recruitment data, which has appeared and grown together with China's first generation of Internet users since the beginning of the millennium, is a piece of data that is almost the same age as the Chinese Internet.

Recruitment Data: Representativeness Problem and Simpson's Paradox

Hiring data is also very difficult data to work with . Only through simple processing, it is difficult to present consistent, representative and valuable information.

The representativeness problem has always been a big problem for recruitment data. What kind of companies recruit online, and what kind of companies choose to recruit from other channels? All along, Internet companies, foreign companies, etc., use recruitment websites much more frequently than state-owned companies and manufacturing companies. This makes the total recruitment volume, total resume delivery volume, and average salary indicators obtained through recruitment data summary have a not small bias from the real national average. The recruitment situation of different recruitment websites also varies greatly.

For example, the following picture is a screenshot of the hot jobs of BOSS Direct Recruitment:

The following picture is a screenshot of the hot search positions for Shanghai recruitment from 58 same cities:

As you can see, the types and directions of recruitment information on the two websites are completely different. When we only use one or several recruitment websites, we inevitably miss a lot, and we cannot output effective conclusions.

In addition, the classification of recruitment data is extremely difficult, which also increases its use threshold. When we use various big data, we often need to match this data with the data of the National Bureau of Statistics according to the appropriate classification standards to get similar caliber data, which is convenient for us to verify the validity of the data.

But for recruitment data, although we have collected 500 million recruitment data and 1.2 billion recruitment vacancies from multiple recruitment website sources through data partners in the past eight years, if we just summarize these recruitment positions, whether it is divided by company, industry or region, it is very difficult to compare with official statistical data.

Why can't massive data get effective conclusions?

First, the number of acquisitions, sources, and company numbers of these data over the years have great differences. As can be seen from the following figure, in 2019, the year with the highest recruitment

number, the total number of recruitment vacancies for all recruitment advertisements nationwide was 340 million, but in 2022 it dropped to 34 million, a full ten times difference.

But the change in the number of recruitments on recruitment websites cannot be completely corresponded to the demand for labor by enterprises. In a booming economy, employees turn faster, there are more businesses, and companies have better expectations for the future. Even the repeated adjustments of the same recruitment information will cause the company's recruitment number to fluctuate more than the real labor demand.

Secondly, the salary of recruitment data is also a mixed variable. The following figure shows the average annual salary from 2015 to 2022. It can be seen that the year with the lowest average recruitment salary is 2017, about 4360 yuan/month. From 2015 to 2017, China experienced a continuous decline in recruitment salaries for two years, and then it rose again.

But did the recruitment salary really fall from 2015 to 2017? Not so.

The first reason for the decline in recruitment salaries is the change in the proportion of social recruitment and fresh graduate recruitment in the recruitment structure. When the website data of the proportion of fresh graduate recruitment increases, the average salary decreases; when the data volume of social recruitment positions with many years of experience increases, the average salary will rise again.

The second reason is the sinking of recruitment websites to second- and third-tier cities, and the penetration of previously less recruited worker groups—for example, blue-collar workers, domestic services and other types of work, in recent years, they have increasingly relied on online recruitment. And the wages of this part of the work are much lower than those of programmers who have mainly recruited through the Internet before, thereby lowering the overall average wage.

The famous Simpson's paradox tells us a result: even if the averages of the two groups are rising, the sum of the averages may decrease. In the following example, when calculated separately, the male admission rate of University A and University B is higher than that of females, but the female with a lower admission

rate is more in the group with a higher admission rate, resulting in a lower male admission rate after the summary.

辛普森悖论

当人们尝试探究两种变量是否具有相关性的时候，会分别对之进行分组研究。然而，在分组比较中都占优势的一方，在总评中有时反而是失势的一方

		申请人数	录取人数	录取率
大学A	♂ 男生	200	100	50%
	♀ 女生	100	40	40%
大学B	♂ 男生	100	80	80%
	♀ 女生	300	225	75%

分开时，大学A和大学B的男生录取率都高于女生录取率

然而，当我们把两所大学的数据汇总时，情况发生了变化

		申请人数	录取人数	录取率
汇总数据	♂ 男生	300	180	60%
	♀ 女生	400	265	66%

汇总后，女生的录取率反而高于男生

城市数据团

Some recruitment websites use their own data to regularly publish salary reports, which are also limited by their website data structure. There are also large differences with the standard industry and occupational structure of the National Bureau of Statistics, making it difficult to compare with other recruitment websites and difficult to combine with statistical data to get a comparable caliber.

Cracking the Simpson's Paradox: How to Group Occupations in a Standard Way

To classify and map the occupations, we first used the GPT4 model of ChatGPT to generate a typical recruitment advertisement.

统性能和可维护性

4. 配合前端工程师，共同完成接口对接工作
5. 负责解决线上系统故障和性能问题，优化代码和数据库

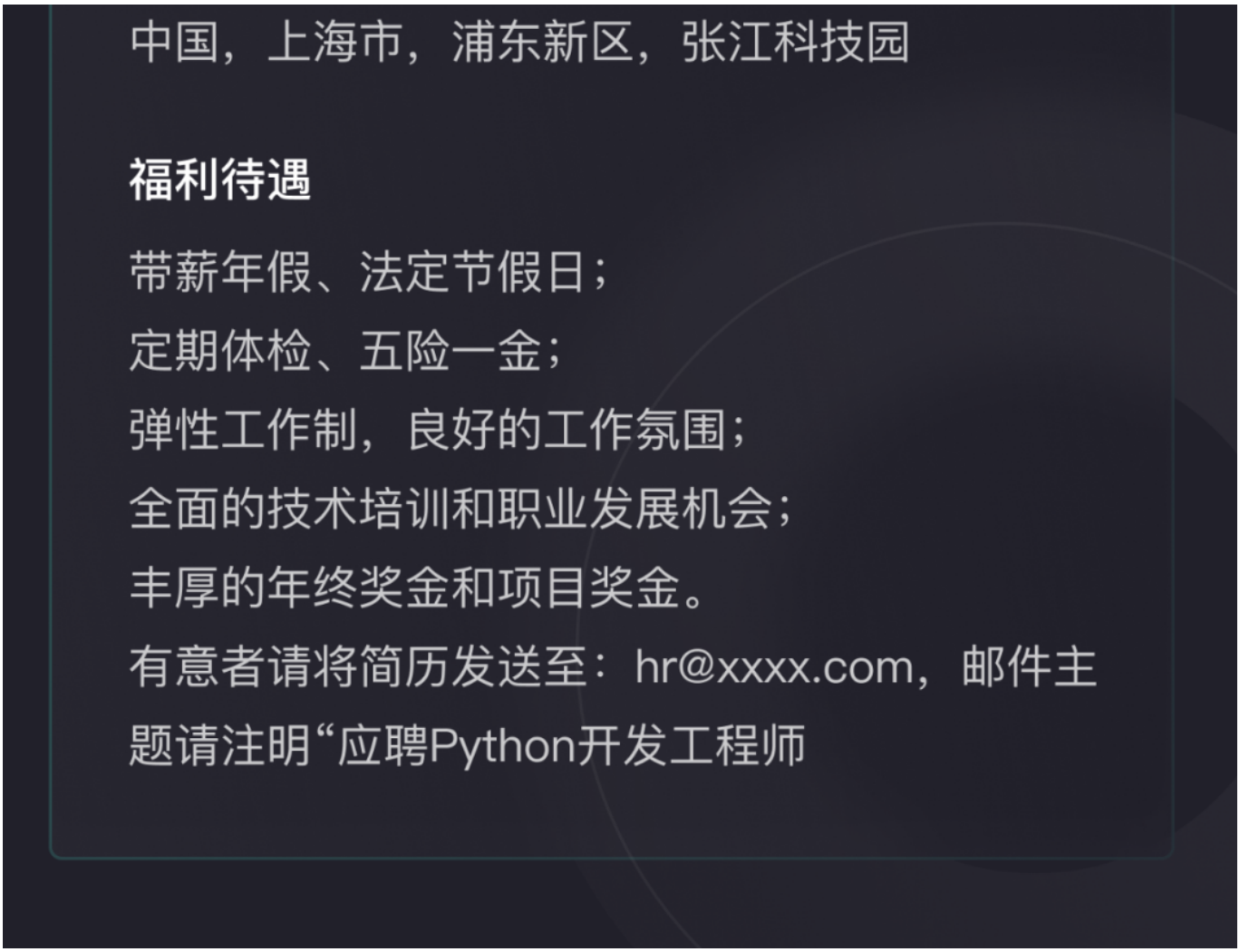
任职要求

1. 本科及以上学历，计算机相关专业
2. 3年以上Python开发经验，熟悉Django、Flask等主流框架
3. 熟悉MySQL、Redis等数据库技术，具备良好的数据库设计能力
4. 熟练掌握Linux操作系统，具备基本的Shell编程能力
5. 具备良好的编码习惯，对代码质量、性能、安全性有较高追求
6. 具备良好的团队合作和沟通能力，能适应快节奏的工作环境

工资待遇

月薪：15,000元 – 25,000元（具体面议，视经验和能力而定）

工作地点



As you can see, the function of the position itself, as well as the required education and experience, have a great relationship with the salary of this position. Education and experience are relatively easy to separate from the text, that is, "computer-related majors, undergraduate, more than 3 years". But how do we classify this occupation? How do we separate a Python engineer from other types of positions in order to control the inherent ability requirements of this position?

The first way is to use the job classification of the recruitment website itself. The following three screenshots are from BOSS Direct Recruitment, Zhaopin Recruitment, and 58 City, and their categories all include "Human Resources/Administration". As you can see, there are many overlaps and differences in their occupational classifications. The "salary performance" of BOSS Direct Recruitment is divided into "salary benefits" and "performance appraisal" in Zhaopin Recruitment. In 58 City, not only are salary and performance combined into one category, but "employee relations" is also included.

When we click into a certain category of occupation, a certain position often "holds multiple positions", or there is only a difference in qualifications, and there is no difference in function. The division of different recruitment data makes the use of recruitment data more difficult.

In order to make a comparison of the same caliber, we naturally need a more authoritative and standard occupational division. We used the "China Occupation Encyclopedia" as the basis for occupational division.

The "China Occupation Encyclopedia" is the occupational classification used by the National Bureau of Statistics, the Ministry of Human Resources and Social Security, etc. when counting various occupations. The China Population Census, Population Dynamic Sampling Survey, etc. all use the "China Occupation Encyclopedia" as the basis for the occupational division of each surveyed worker. The latest 2022 version of the "China Occupation Encyclopedia" includes 8 major categories, 79 medium categories, 449 small

categories, and 1639 detailed categories (occupations), which is the most complete and authoritative division of Chinese occupations.

For example, if we want to find the classification of "programmer" from it, we can find it through the following hierarchical table:

This kind of occupational division ensures that the intersection between occupations is minimal and the union is maximal to the greatest extent. We will try to map all recruitment data to these 1639 occupations.

But how to divide and map? Just the position of "computer program designer", the job title on the recruitment website may include a series of keywords such as JAVA, Python, Ruby, Golang, Node.js, C++... This is a relatively familiar occupation for the author, and we may be able to traverse this category of occupations through keyword mapping. But for some relatively unfamiliar occupations, such as "course consultant with a monthly income of over 10,000 and a large space for promotion", can you accurately classify him into the category of "salesperson" in the standard occupational code?

Therefore, we used a method of text learning. First, let the computer learn the specific work of each occupation, and then match it through the job description of each position, as shown in the figure below:

1. 标注职位：

Python开发工程师

4-04-05-01 计算机程序设计员

2. 汇总被标注职位的职位需求：

Python开发工程师

- 负责公司产品后台服务的设计、开发和维护；
- 根据项目需求，进行模块设计、编码、测试等相关工作；
- 参与产品架构讨论，改进现有系统，提高系统性能和可维护性；
- 配合前端工程师，共同完成接口对接工作；
- 负责解决线上系统故障和性能问题，优化代码和数据库。

……

3. 根据被标注职位与职位需求的上下文关系，学习各类职业在职能中的词语向量，计算各词语之间的向量夹角（词对概率）：

- Python开发工程师（计算机程序设计员）——模块设计，词对出现概率0.8
- 课程顾问月入过万（营销员）——与客户保持良好沟通，词对出现概率0.7
- 钳工包吃住（机修钳工）——拆卸检查机械设备，词对出现概率0.9

4. 对于所有未分类职业，根据其职能要求映射其行业

负责公司产品的核心模块和功能的设计、开发、调试和维护（ A ）

A.计算机程序设计员，概率0.93

B.根营销员，概率0.05

C.机修钳工，概率0.02

城市数据团

Through preliminary annotation, the specific work of each occupation is combined with the name of the occupation to calculate the high-frequency word pairs of the occupation-function. Starting from the job function of the recruitment advertisement description, use Bayesian probability to calculate which specific occupation may correspond to, and calculate the specific classification of each occupation like a cloze.

cocos2dxvc程序员可培训	网游JAVA程序员	python程序员金融量化投资	移动web开发
PHP程序员职位编号10	前端程序员开发学徒理	UI设计web	网站程序员
c语言测试web软件开发	web全栈开发	长沙中高级web前端工程	web前端开发主管六
AspnetWEB程序员	中高级web前端开发上市	web前端实习生提供食宿	ERP程序员TIPTOP
0基础web前端五险	汽车行业web前端开发助	初级PHP开发程序员五险	web前端开发3500
机加程序员	PHP程序员网页编程网站	丰富java程序员诚聘	IT程序员20届软件工程
网站程序员phpdedecms	数据库程序员DBDP	web前端开发实习生饭补	核心程序员软件开发工
初级WEB程序员	php程序员中高级别均可	无人机系统程序员	程序员15天休

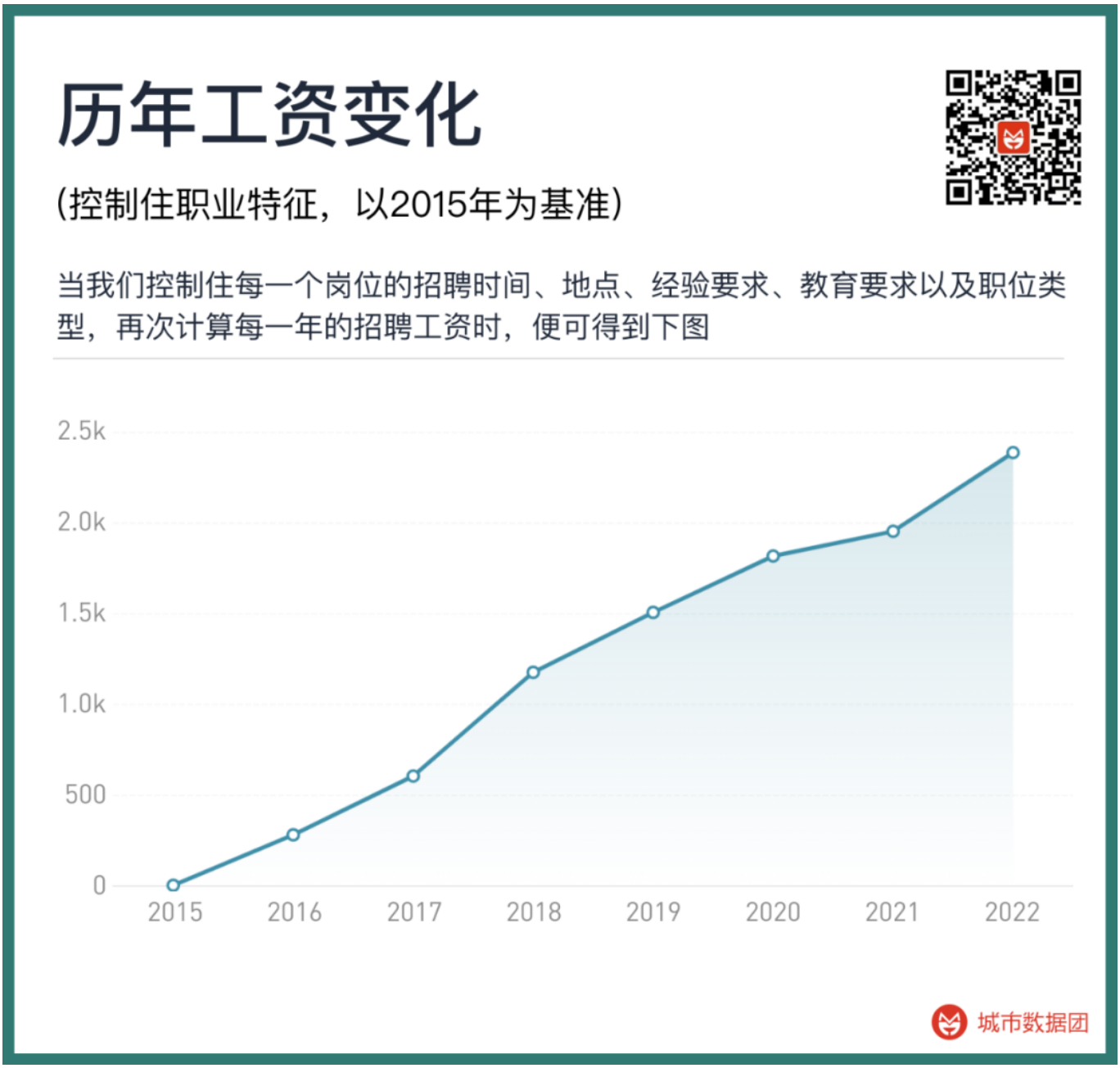
This method has a very high accuracy. The following is a group of examples we classified into "computer program designers". As you can see, even if there is no keyword "programmer" in the title of the position, we may not be able to traverse various program-related keywords, but we can accurately classify this position through its job function.

Using this method, we have allocated the recruitment data of 500 million and 1.2 billion recruitment vacancies obtained from various recruitment websites, as many as 18 million occupations, to more than 1,500 standard occupations, forming a standard occupation database of various cities across the country from 2015 to 2022.

Initial Glimpse of Occupational Passwords, Recruitment Data "Mine" Built

With the standard occupational database, we can control the recruitment time, place, experience requirements, education requirements, and job type of each position. When we use these data again, there will be no bias problems similar to Simpson's paradox.

For example, based on this, when we calculate the recruitment salary for each year again, we can get the following figure:



At this time, we can see a steadily rising wage growth curve, without a sudden drop. The same recruitment time, place, experience requirements, education requirements, and a job, the recruitment salary in 2022 is 2385 yuan higher than in 2015.

This also means that this "gold mine" containing billions of recruitment data is no longer a chaotic open-pit wild mine, but has been built into a stable and controllable industrial-grade mine.

What needs to be done next is to mine and smelt various valuable gold information from it.

Preview of the next part

In this post, we figured out how to map the positions on job advertisements to more standard occupational classifications, so as to avoid data misjudgment caused by incorrect classifications. This just completes the construction of the "recruitment data mine".

In the second post of the series, we will enter the main part of mining (data analysis) to see how the income of practitioners in various cities and occupational categories changes with time and experience in their

careers Cumulative improvement and change. For readers who want to do it themselves, we will also simultaneously update the query module of all research-related data in this series (including the AI replacement rate that everyone cares about.） The calculation of the AI replacement rate for different occupations will be introduced in detail in the third post.