

***CITY UNIVERSITY OF HONGKONG***

**COM 5507 – Social Media Data Acquisition and Processing**

**Opportunity and Challenge of Operators in Big Data Era**

**——Visualization and Analysis of Operator’s information in Bosszhipin**

**CAI YANHUAN 55474896**

**HE WEIYUAN 55338942**

**ZHANG CHUHAN 55467217**

**CHEN DANTONG 40089057**

**Instructor: ZHANG XINZHI**

**December 12, 2018**

1. **Background**

Nowadays, data has become increasingly important in many fields. For many enterprises, using big data to better optimize their business had been the icing on the cake. Besides, with the development of big data technologies, enterprises start to use data system to satisfy their goal to accurately target consumer as well as to provide preferable user experience and better products and service. Thus, data-related positions whose responsibility is systematically and efficiently processing data are extremely in need. If we take a look at a quite common position in recent years, operators, a trend will be found in this area. The number of operator positions which require job candidates to be equipped with data analyzing and processing skills are continuously climbing up. A well-known venture capitalist and technology executive LI Kaifu has predicted that in the future many positions will be replaced by artificial intelligence(AI). Positions such as operators will be the first to bear the brunt. In this case, what should operators do faced with the possible risk and will there be a transition in this area?

1. **Literature Review**

We found several studies about the job analyst. In 2007, Zoe\_Achelics used python to scrape information on data analyst in her article “*互联网企业数据分析职位招聘现状分析 ——以拉勾网为例*". She used various of graphs to present several analyst perspectives including position demand in different cities, responsibility requirement and salary distribution, company type and salary distribution, company development stage and salary distribution, the responsibilities of data analyst, and data skills of data analysts.

The second article we consulted was “*数据分析师就业分析报告*". This report gave a general description of the recruitment of the data analyst position. It concluded that most of the position distributed in the first-line city and Hangzhou. The salary of the position mainly concentrated on six thousand to twenty-six thousand yuan. Besides, the position mainly recruited young people, requiring one-to-three year working experience. Among the positions, those requiring more working experience also offered a higher salary.

Moreover, in the article “*Kaggle调研数据揭秘如何成为一名合格的“数据科学从业者?*”, the author provided a reference for analyzing the position of data science-related practitioners. The author found out the relationship between high salary and company type as well as the practitioner's major. Besides, he also indicated the data analysis tools that data field personnel uses most. After reviewing these articles, we decided on the study of the challenge and opportunities of operators in current Chinese marketing.

1. **Research Question**

Through our data, we want to explore the marketing requirements for operators in China and make a comparison of differences of operators' requirements and required quantity among four cities. Our research questions are listed below:

1. Which city will offer the highest salary for operators?
2. Which kind of operators will be highest in demand?
3. How many years of operation working experience are in higher demand? & Is longer operation working experience related to higher salary?
4. Which level of education are most required for operators?
5. Of which company types the demand for operations is highest?
6. Which kind of abilities the do company want most?
7. How many positions demand the ability of data analyzing? Is the “data” ability really related to higher salary?
8. Which factors will affect the average salary most significantly?
9. **Data Collection**

The first step of data collection was to choose the target website. We found several Chinese mainstream job recruitment websites such as Lagou, BossZhipin, 51job, Zhilian. After trying out scraping data from these websites, we chose BossZhipin as our target website. There are some reasons why we chose this website: compared to 51job.com, the structure of this website is clean and tidy; there were enough data for us to analyze in BossZhipin, and their anti-scraping mechanism was not so difficult as Lagou.

The next step here is to clarify the structure and information we need in BossZhipin. We found the information we need is concentrated on two kinds of pages: the category page and the detail page, which are shown in the pictures below.



*(Category Page: This page includes position, city, salary range, company, company size, company type, required experience, education level and links of detail pages )*



*(Detail Page: This page includes position, city, salary range, company, company size, company type, required experience, education level and responsibilities)*

Then, we decided on our scraping strategy. As shown in these two pictures, we could see that most of the information included in the first page is also included in the second detail page. So our strategy of scraping could be divided into two steps. The first one was to scrape all the links of detail pages from category pages, then to scrape contents needed from the links we scraped in the former step.

Next, we could carry out data collection according to our strategy. We used beautifulsoup and XPath libraries and defined seven python functions to scrape the information and store information in the format of CSV. In order to use these tools, we need to start from imports. (Note: make sure that you have installed lxml, requests, and BeautifulSoup.)



We found some regulations of links, so we defined a function to create URLs of category pages and used BeautifulSoup to get contents of these category pages straightforward. 

After that, we scraped the links of detail pages from contents of category pages and aggregated all the information into one list.



Then we can executed the second step of our strategy, which was to scrape specific information from all the links of detail pages.

Take one of the detail pages as an example. A function was defined to send a request so that we could change the request headers easily. And we defined another function to scrape the specific information(including publish-time, position, salary, company name, company type, company size, responsibility, and requirement), which used XPath language.



After scraping, we needed to encode all the information with utf-8 and store them in the format of CSV. In this last process, we defined three functions, including saving information and saving box header and the main function which connects all the function we defined before.



The most difficult problems we faced here is the anti-scraping system in BossZhipin. In this website, it set double anti-scraping mechanism. One of them was verification code. When the website detected that the access frequency was too high, a verification code popped up for us to enter, confirming that the visitor was not a robot. Because of that, our scraping process was always interrupted because of the verification code.

The other one of the anti-scraping mechanism was the 403 forbidden error. Due to the high frequency of requests, the website would block your requests from python and display 403 error in the page, which showed that users didn't have permission to access on this server. The users usually were blocked within 24 hours in Bosszhipin. This question was harder than the verification code. Cause if we were blocked, we only could use our header to scrape data after 24 hours.



With regard to verification code, we used manual operators to avoid anti-scraping of the website. We could access the verification code manually when the programme is interrupted by this system. While about the 403 error, we changed our User-agent. In the Requests library, the user is allowed to customize the request header information, so we changed the value of the User-Agent in the request header so that the server could be spoofed to achieve the purpose of bypassing the anti-crawler.



In the whole scraping process, we used several User-agent to fake our request header.

Mozilla/5.0 (Windows; U; Windows NT 5.1) Gecko/20070309 Firefox/2.0.0.3

Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_14\_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/70.0.3538.110 Safari/537.36

Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/70.0.3538.77 Safari/537.36

Mozilla/5.0 (Xll; Linux x86\_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.89 Safari/537.36

Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/70.0.3538.77 Safari/537.36

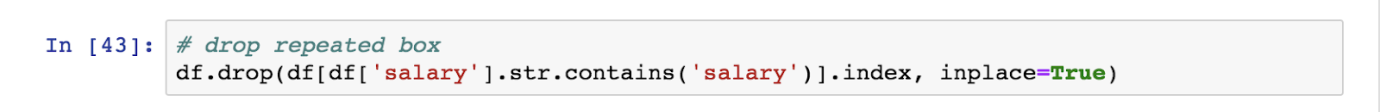
1. **Data Cleaning**

At last, a total of 1227 pieces of raw data were acquired finally and stored in CSV. Separated CSV files were imported into a Jupyter notebook and merged into one data file.



*(Raw Data)*

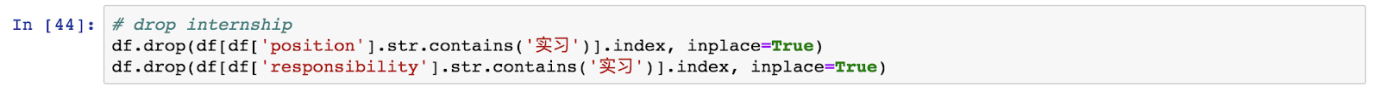
First, we filtered the repeated box header and repeated data initially because it would influence the cleaning of purely numeric data(salary). We dropped repeated data by judging if there are repeated data of the publish time, position and companyName.





Then, the items which are internship jobs were deleted, and 1168 pieces of data were obtained.

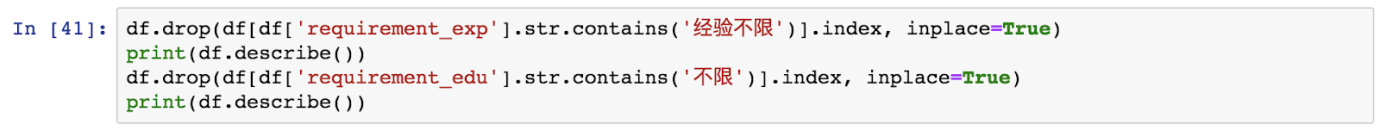
According to our observation, internship jobs recruitment information would clarify the internship situation through the statement of position and responsibility.



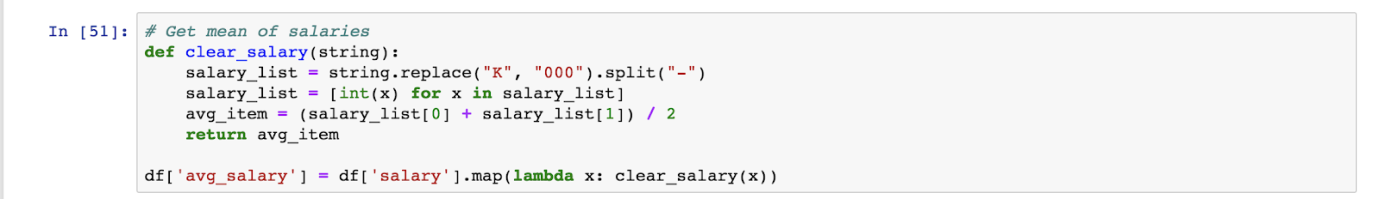
Next, since there are three different kinds of information in the requirement column, which includes city, experience, and education, we define a function to split the three items in ‘requirement’ into three separated columns.

As the code shows, we use to define a division function which has two arguments, string, and index. Firstly, we used ‘:’ to split the string. Secondly, we used the index to help us to output which information we want. For example, we wanted to divide the city in the string. We assigned l[index] to ret, which meant that we could get the second element in "l" list. Then if the index is not equal to 3(this condition is set for scraping the education level), we use ret[:-2] to get "北京” from “北京经验”.  In this example, we saw every word as an element in the string.

We delete the items whose education requirements and experience are unlimited, which will not show on our data analysis.



As for salary here, we could see that the data in the salary column was a range instead of single numeric, which was not suitable for analyzing. So we decided to get a mean salary of the range. We used "-" to split this range, replaced  “K” with “000” and changed the string to numeric for the convenience of calculation. After that, we calculated the average salary and used the map to add ‘avg\_salary’ to our data frame.



After the cleaning steps above, consistent data(screenshot are as follows) was got. Then we can enter the next visualization stage.

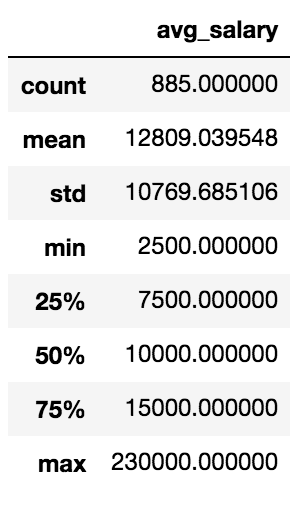
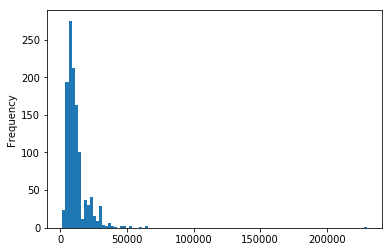


1. **Data Visualization**

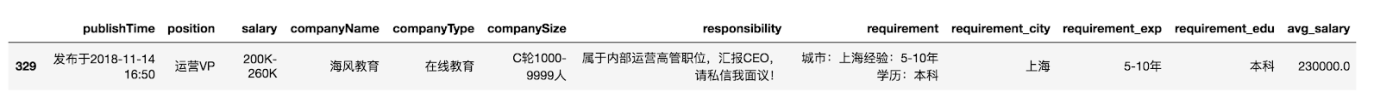
We will start from importing consistent data here. There are 885 consistent operator recruitment records here for us to analyze.

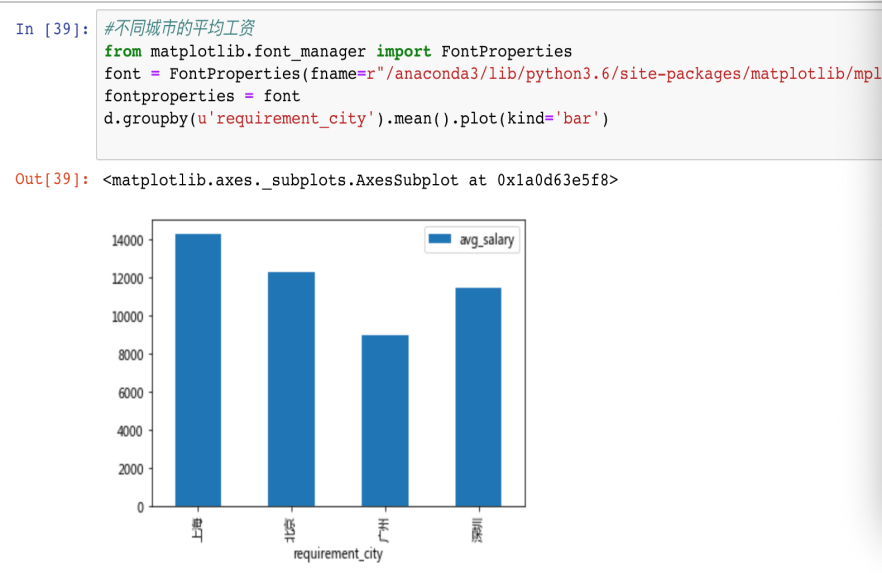
**Q1. Which city will offer the highest salary for operators?**

The average salary of all four first-tier cities(Beijing, Shanghai, Guangzhou, Shenzhen) is 12,809 yuan, the maximum is 230,000 yuan while the minimum salary of operators is 2,500 yuan. From the histogram below, the distribution of salary is significantly right-skewed. It suits the marketing regulations. Few people can get significantly high salary.

The maximum salary is for an operation VP, which reaches 230,000 yuan a month.

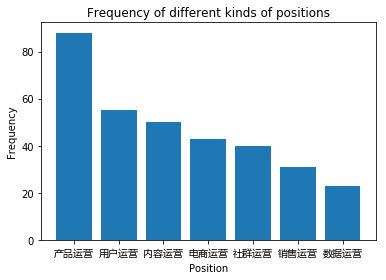




From the bar chart above, it can be seen that Shanghai will offer the highest average salary for operation talents and Guangzhou will offer the relatively lowest average salary for operators. After combining observation and economic status of these four cities, it is reasonable to display in this way. In Guangzhou, the living consumption and house rent are relatively low compared to other first-tier cities.

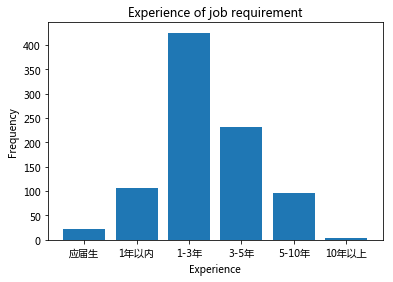
**Q2. Which kind of operators will be highest in demand?**

As we can see on the graph below, product operators are highest in demand. The user operators, content operators, e-commerce operators, sales operators, community operators, sales operators, and data operators are listed in descending order.

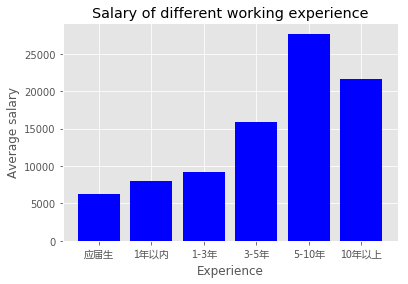
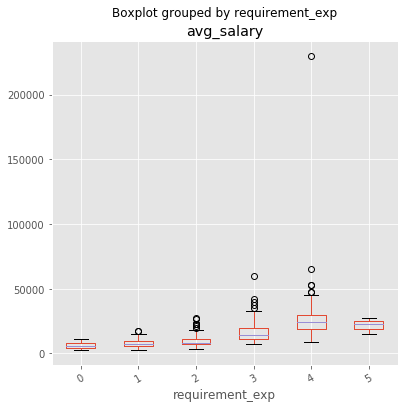


**Q3. How many years of operation working experience are in higher demand? & Is longer operation working experience related to higher salary?**

As the graph shows below, most positions need 1 to 3 years’ experience, 3 to 5 years is the following one. Only 3 positions require above 10 years' working experience, accounting for 0.25% of the dataset. The fresh graduates are the second least-demand, which indicates that the position of operators needs applicants with some working experience. Also, the employee with above 5 years' experience is in low-demand shows that the industry is relatively young, few people persist more than five years.

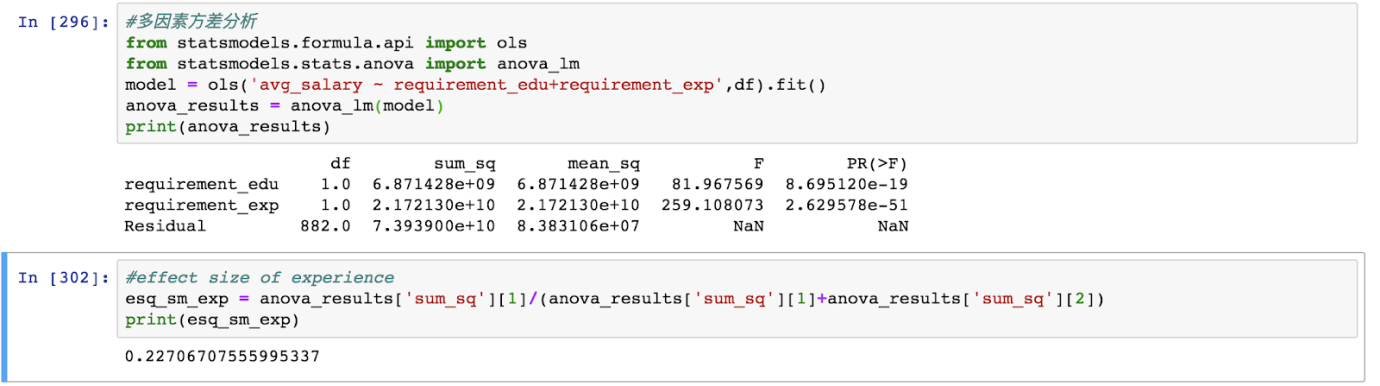


Speaking of working experience, we can see if longer operation working experience means a higher salary. From the bar chart, We can see that the average of salaries of operators within 3 years is generally below 10000. While operators with over 3 years working experience can try to ask for more than 15000 one month. When operators work in this field for over 5 years, they can gain over 20000 per month. The salary can jump greatly in different stages.

In the box plot, there are several discrete numeric values in the “3-5 years” and “5-10 years” variables. We are not pretty sure about the relationship between the experience and salary from the box plot.

After recoding the experience, we use ANOVA to test our assumption. There was a significant effect of working experience on the salary at the p<.05 level.

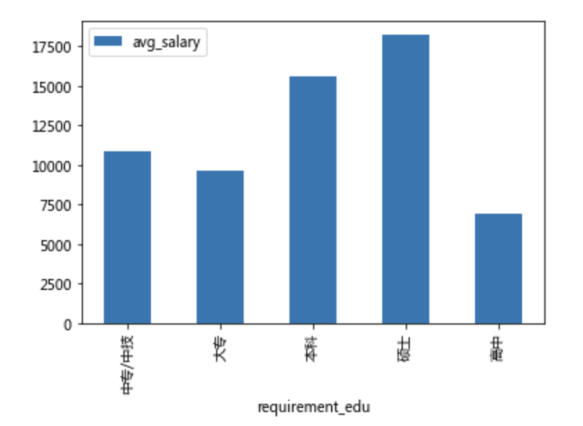
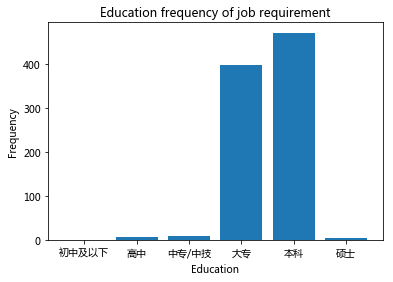


According to the responsibilities of these positions, we can find that all the position which demand over 5 years refers to management responsibility. Overall, almost all the operations will transfer to the manager to add value for themselves.

**Q4. Which level of education are most required for operators?**

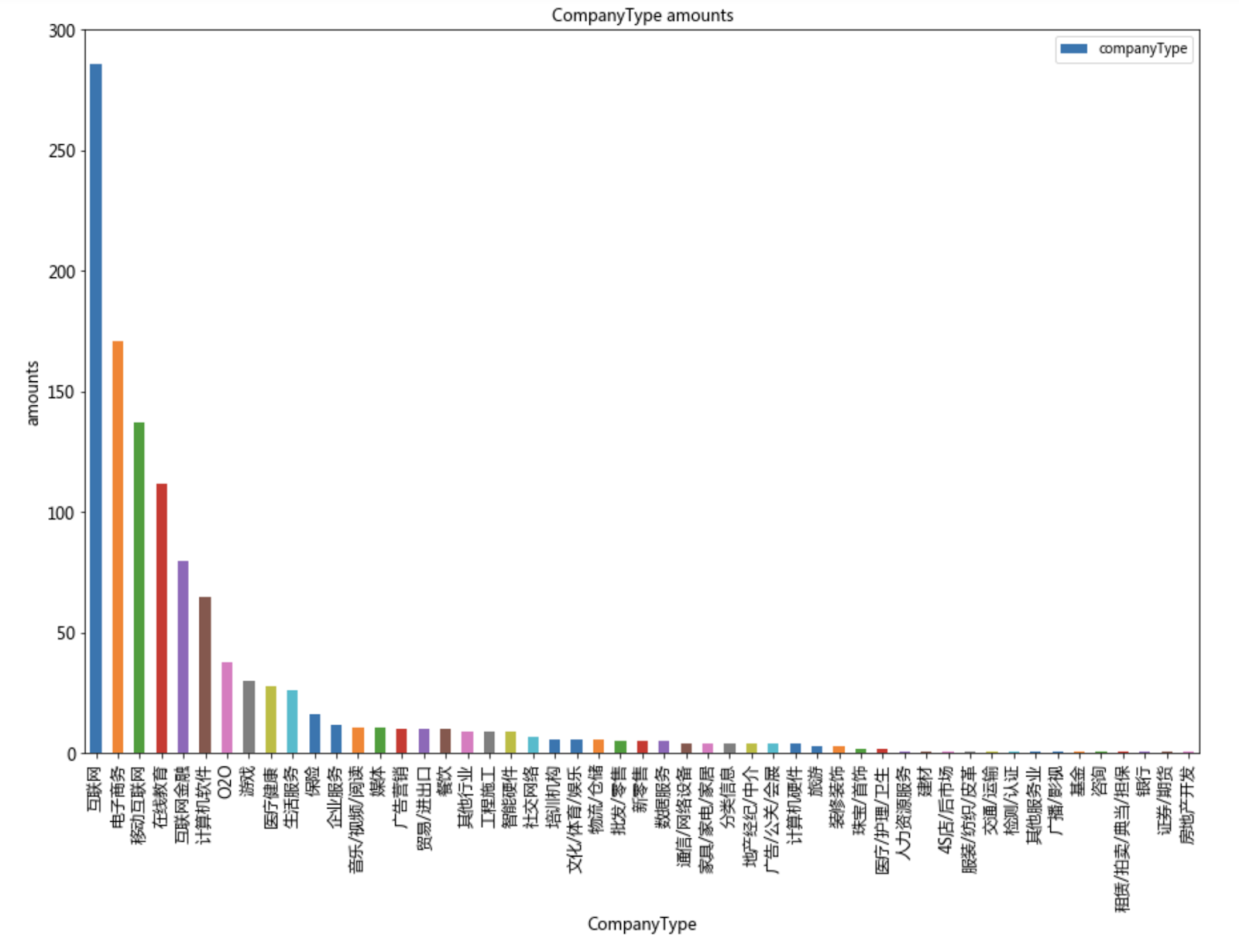
The most welcome academic degree is a bachelor's degree. What comes next is associate degree. The position of operators somewhat needs people who are highly-educated. But bachelor's degree is enough for this position since the master's degree is not so needed to be indicated by the bar chart.

The average salary ascends with education degree except for the associate degree whose average salary is unexpectedly lower than the technical school graduates’.

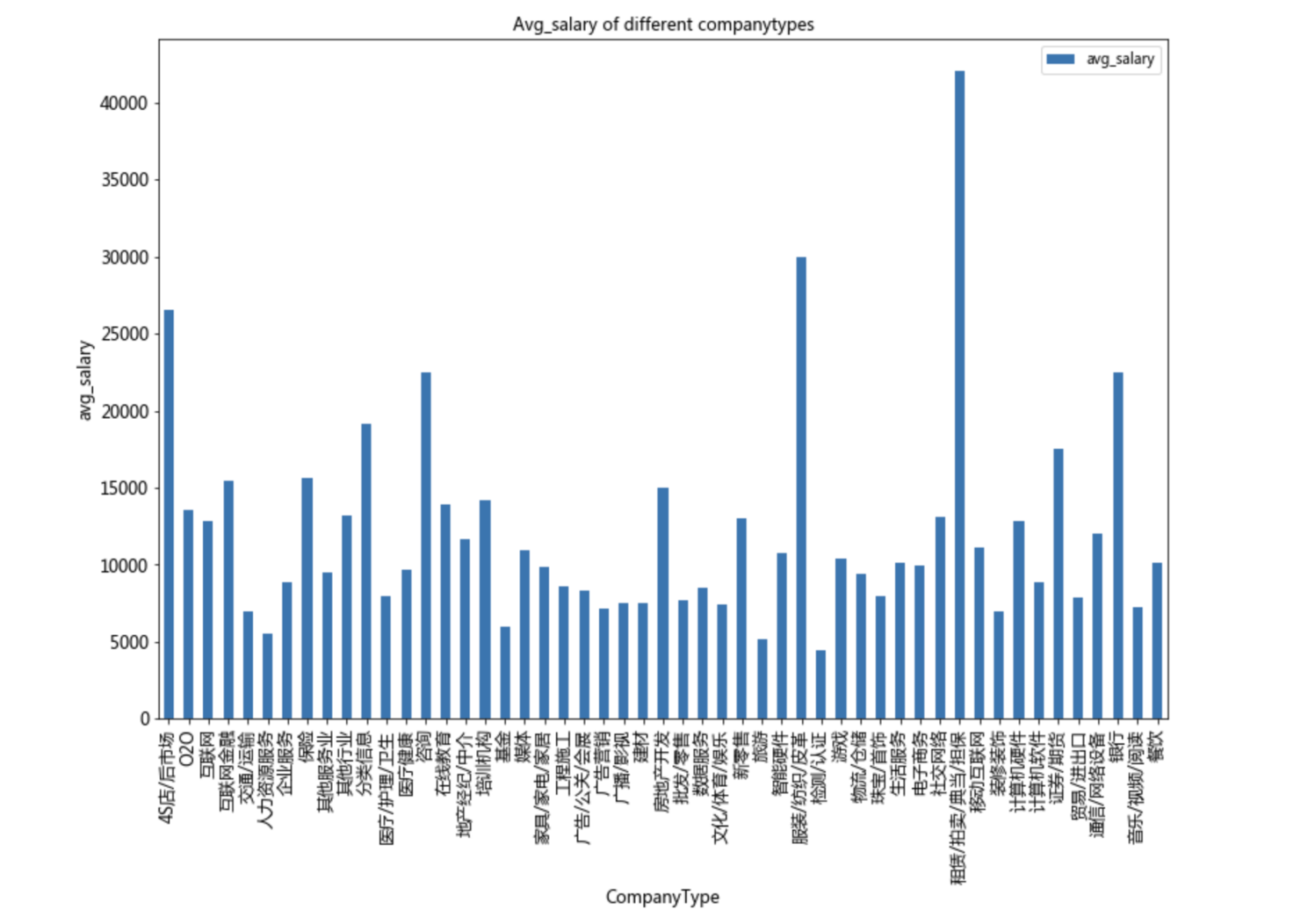


**Q5.  Of which company types the demand for operations is highest?**

Among the 51 types of company, the Internet industry is the largest employer of the operator, far more than other company types with nearly 280 positions of the operator released by the Internet company. Electronic commerce, mobile internet, online education, i-finance, and software come after.

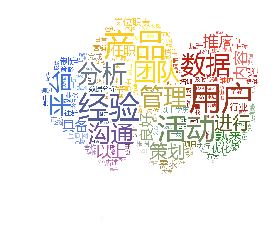


The average salary in the different types of company differs from one another. Excluding the outlier, the only one demand in the pawn industry, the average salary is relatively high in the garment and textile industry, 4S store, consulting company, and bank.



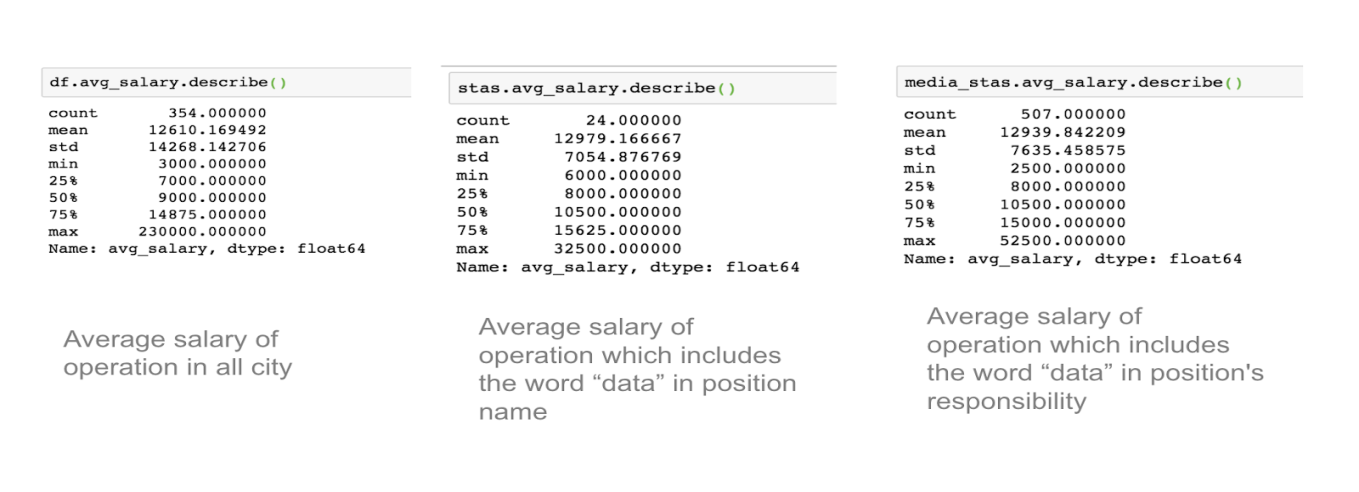
**Q6. Which kind of abilities the do company want most?**

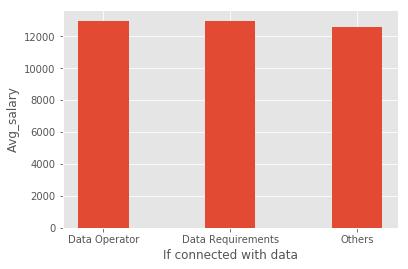
After analyzing the word frequency in the description of responsibility, we found the words USER, EXPERIENCE and PRODUCT are the most frequently occurring words in responsibility. Also, the words ACTIVITY, DATA, and ANALYSE appear quite frequently.



**Q7. How many positions demand the ability of data analyzing?  Is the “data” ability really related to higher salary?**

Fromthe data shown below, we can see that there are around 24 positions named data operator and there are 507 records mentioned "data" in the responsibility of the position, which means around 60% of whole operators we scrape are related to data. And the average salary of positions is higher as the relationship with data gets closer ( from requiring data analyzing to data operators).

****

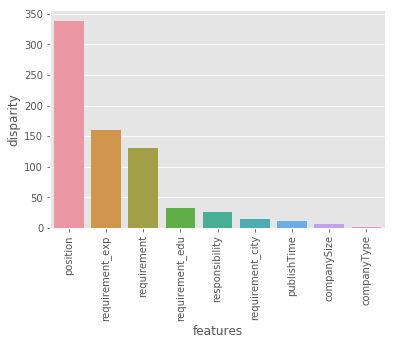
****

The difference of data operators and operators whose requirements includes data is mainly their responsibilities. The data operators will focus on data processing and analyzing in order to provide support for the side of business, most of whom are other operators. The operators whose requirements include data will emphasis more on their own abilities, for example, user operators will manage the user policy and increase user viscosity. At the same time, they are required to have basic data analysis skill, such as Excel and mysql.

**Q8. Which factors will affect the average salary most significantly?**

We define a function which calculates p values. From those p values, we'll calculate a disparity score. Higher the disparity score, better the variable in explaining the salary.

From the graph, we can see that position turned out to be the most important factor followed by working experience, education level, responsibility, and working city.

****

1. **Conclusion:**

Viewing the result above, we can find some interesting relation between Internet and operators. For the name of the positions, there are product operator, user operator and community operator, which are closely related to the Internet. Also, there are Tianmao operator, Amazon operator, and e-business operator, which are totally relied on the Internet. In turns of the industry, apart from the Internet industry, which ranks first in the amount of position, there are many other industries related with the Internet such as E-business, Internet finance, online education, social network, and so on. As a position to maintain the daily content, activities and so on, it is actually growing as the Internet industry grows. In traditional industry, like manufacturing and commerce, content and activity are not the core value of the industry. However, the information technology brought new chances and needs, leading to the development of an operator position.

From the data we collected, we find some similarity of the position and the Internet industry. For instance, the Internet industry is comparatively young, and working experience requirement of the operator is quite short as the largest part is 1-3 year experience. The operator is a position requires comprehensive ability and especially the ability to absorb new knowledge and acquire new skills for it needs the continuous catching up with their product, users and social environment. The ability of learning and adoption is more important than long experience and high education degree. Similar to other jobs, the salary of operators is positively correlated with an education degree and working experience, while there are not a large number of positions requiring high education level(master or higher) and long-time working experience(more than 5 years).

What’s more, as the usage of big data come to the stage, it goes further with the help of data analysis. The operator positions require the ability of data analysis and even some of their name is data operators. Since big data is increasingly important in this age, the ability to use it in operators is very useful. Among the positions, the more closely related to data analysis, the higher the salary is.

For people who head to this job. We have some suggestions. Firstly, to be a good operator, one have better to master the skills of data analysis and data management. As many of the company do not clearly understand the value and application of big data, usually there is not clear Instruction for the operator to manage the data. With the skill, operators can make better use of the data and accumulate this resource. Besides, for an undergraduate student, to get an operator job, working experience is more important than a higher education degree. As most of the companies require at least one-year working experience, applicants with relative internship are more competitive. Last but not least, to be an operator, applicant have better to plan the working career earlier and be ready for transformation. As this position usually require less specific skill and experience while repeating the same contents, the operators might apply their accumulation in some more challenging job such as product manager.

1. **Implication**

For the further study on some related topics, it can provoke to explore more about the challenge and opportunity in an era of big data. We can crawl more data to show the whole marketing situations not only from one position. And also we can compare the different responsibilities of different positions to clarify what kind of talent will be more competitive and popular in today’s world.

For the job seeker who wants to engage in the position of operators and obtain a high salary, this report would be a reference for him/her to select related jobs and prepare the required skills and other factor needed in advanced. For example, they would better equip themselves with data processing knowledge and skills since they are faced with the challenge of big data.

For our future study, the process of finishing this report will play a role of enlightenment. It improves our capability of coding, the skills of searching, and most importantly, change our way of thinking. Besides, although we spend a lot of time and make enormous efforts in python crawler, there is something imperfect when presenting the result. We should have connected each separated result together. So this would be a point to carefully think about when we or other data learners want to do some similar scraping —— to consider how to tell a story before data acquisition. In this way, data can serve the theme better.

**References**

Zoe, A. (2017). 互联网企业数据分析职位招聘现状分析 ——以拉勾网为例. Retrieved from <https://www.jianshu.com/p/c0fc94df495d>

Lao, C. (2018). 数据分析师就业分析报告. Retrieved from <https://zhuanlan.zhihu.com/p/39337971>

RS, J. (2018). Kaggle调研数据揭秘如何成为一名合格的“数据科学从业者”？. Retrieved from [https://mp.weixin.qq.com/s?\_\_biz=MzI1NjY2NzUxOQ==&mid=2247487615&idx=1&sn=887a5aab632fba991c40675ef2f1b5ef&chksm=ea226a6fdd55e379c65be19f29163863e9ee2157d21103895424e777d763786640b9d154dfb1&mpshare=1&scene=1&srcid=1210SnKMqGFxTxwBXYqEHR5V#rd](https://mp.weixin.qq.com/s?__biz=MzI1NjY2NzUxOQ==&mid=2247487615&idx=1&sn=887a5aab632fba991c40675ef2f1b5ef&chksm=ea226a6fdd55e379c65be19f29163863e9ee2157d21103895424e777d763786640b9d154dfb1&mpshare=1&scene=1&srcid=1210SnKMqGFxTxwBXYqEHR5V" \l "rd)