**Workload Distribution**

**For data acquisition and processing:**

Data acquisition-LI Weijia

Data cleaning-Mao Xianghua, LI Weijia

Data analysis-LI Weijia, Mao Xianghua, Zhang Xiaoguang, Huang Huanhuan

**For presentation:**

Introduction-Zhang Xiaoguang

Data cleaning-Mao Xianghua

Data analysis- LI Weijia, Huang Huanhuan

**For final report:**

Abstract& Introduction- Zhang Xiaoguang

Data acquisition-LI Weijia

Data cleaning-Mao Xianghua

Data processing& Results& Conclusions-Huang Huanhuan

Edit, organize and submit paper-LI Weijia

**Data-oriented Employment Advice For The Graduates of Department of Communication and Media, City University of Hong Kong**

Goup member

|  |  |
| --- | --- |
| LI Weijia | 56924596 |
| MAO, Xianghua | 56790971 |
| ZHANG Xiaoguang | 56832260 |
| Huang Huanhuan | 56316413 |

COM5507: Social Media Data Acquisition and Processing

Dr. LIU Xiaofan

December 10, 2021

**Data-oriented Employment Advice For The Graduates of Department of Communication and Media, City University of Hong Kong**

# 1. Abstract

Job seeking is important to everyone who is going to graduate next semester. However, the information gap between the college and industry still exists, that's why many students feel confused when facing the challenge of looking for a suitable job. In this study, we analyzed the occupations that previous graduates of the Department of Media and Communication, City University of Hong Kong usually worked on, including advertising, public relation, product manager, business analysis, user research, and content operation, that are introduced by the official website of the department.

We chose the website "Shixiseng" as the object of data crawling, and used BeautifulSoup, regular expressions, pandas, etc. to crawl, clean, analyze and visualize the data. We analyzed the distribution of each position, the salary of each job in different cities, the salary of each job in different companies. Finally, through data analysis, we found that there are significant differences in salaries between different positions.

# 2. Introduction

## 2.1 Purpose

Students in the Department of Media and Communication will soon be faced with the challenge of finding a suitable job. This is very important for those who never worked as a full-time employee before. Most students choose this major and will work in related industries in the future. And they hope the future job will meet the professional training in college and get an ideal salary. However, many students majoring in media and communication may feel confused when faced with so many choices, given the information gap between the industrial field and academic field. So, we’d like to explore the information about the job our graduates usually take on.

## 2.2 Meaning

According to the “Graduate Employability Research Report 2021” released by Xinhua Net (2021), college students pay more attention to social reality when choosing a career. Salary is the most weighted indicator, followed by job stability and settlement issue of one city. Nearly a quarter of students are eager to get a job from Internet companies with relatively high salary levels. Based on this report and the website we choose to crawl, we can infer that when choosing a career, salary, career direction and city are the three major indicators, and we should analyze them in our study. By analyzing these three indicators, we tend to give a reference to our graduates.

Specifically, as for the career directions in the media and communication field, we adopted the former graduates’ occupation choices introduced by the department of communication and media’s progames, and classify them into these jobs: advertising, public relation, product manager, business analysis, user research and content operation.

Based on the three indicators and occupations mentioned before, we should look at where to find these jobs, thus, we put forward the first research question:

**RQ1: For each job, which cities provide the largest nubmber of positions? (For number > 10)**

According to the public statistics, even inside the richest four cities in mainland China, the average salaries still have gaps (Yicai, 2021). We assumed that, for each job, the salary is not always the same, so the second research question is:

**RQ2: For cities in RQ1, how much salarise they provide for corresponding job?**

Similarly, for the same job, different companies may provide different levels. The third research question is:

**RQ3: For each job, which companies provide the most salaries?**

As mentioned before, many students hope to be enrolled into the Internet companies because of the salary. After comparing job in different companies and cities, we’d like to know, among different positions, is it real that the salaries differ a lot? So here is our fourth research question:

**RQ4: Is there any significant difference in salary among different positions?**

# 3. Method

## 3.1 Data acquisition

***Data source***

Our data source comes from "Shixiseng". It’s a campus recruitment platform for school recruitment, provides college students with more than 250,000 corporate internships and campus recruitment information including domestic and foreign industry giants. Therefore, we think it will be helpful for us students in school to find a satisfying job.​

***Variables***

The base URL is "https://resume.shixiseng.com/interns?Page=", and we used for loop to get ten pages for each post, which means that 200 pieces of data were crawled for each post. The job title, company name, workplace, salary, identity requirements, academic requirements, company scale, company introduction, and details page URL from this website were scraped.

**Figure 1**

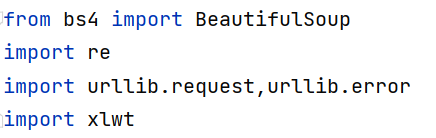


***Procedure***

We used four packages including BeautifulSoup, re, request and xlwt to get data.

* BeautifulSoup is a Python library that can extract data from HTML or XML files. It can use your favorite converter to achieve idiomatic document navigation, search, and modify documents. Beautiful Soup will save you hours or even days of work time.
* The re module makes the Python language have all the regular expression functions. The compile function generates a regular expression object based on a pattern string and optional flag parameters. This object has a series of methods for regular expression matching and replacement. The re module also provides functions that are exactly the same as these methods. These functions use a pattern string as their first parameter.
* The get() method in the requests library can send a request to the server. The request type is the GET method of the HTTP protocol; the post() method can also send a request to the server. The request type is the POST method of the HTTP protocol. Depends on the webpage visited.
* "xlwt" is a library for developers to use to generate spreadsheet files compatible with Microsoft Excel versions 95 to 2003.

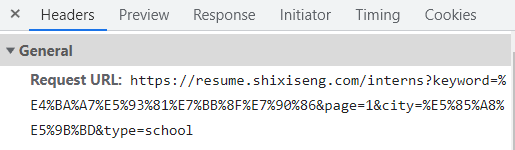
**Figure 2**



The major difficulties to scraping the pages are to analyze the URLs structure, use regular expressions to match and get exact results, and the access to the web page was denied after several tries.

* For the first problem, we found pressing the F12 key on the website to do the inspecting, going to the Network, and to the Headers, to the General, and finally, we can get the Request URL we want to analyze. "keyword=" follows the search keyword, "page=" follows the page of the posts, and "city=" follows the working place.

**Figure 3**



* The second problem is the most time-consuming part during scraping. We tried dozens of times to revise the regular expressions to get the correct results. After figuring out the structure of the website, we used nine regular expressions to get the nine variables we wanted to analyze for each post. Following are those regular expressions used to scrape:

findtitle=re.compile(r'<a.\*?>(.\*?)</a>')

findlink=re.compile(r"(?<=href=\").+?(?=\")|(?<=href=\').+?(?=\')")

findintro=re.compile(r'<span.\*title=.\*>(.\*)</span>')

findlocation=re.compile(r'<span class="city ellipsis" data-v-d5abf57a="">(.\*?)</span>')

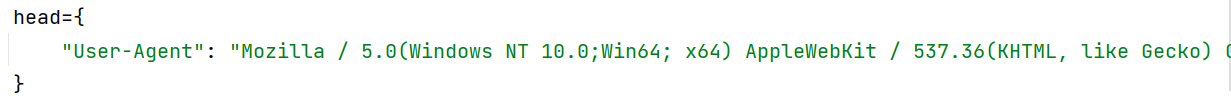
findproperty=re.compile(r'<span class="ellipsis" data-v-d5abf57a="">(.\*?)</span>')

findsalary=re.compile(r'<span class="day font" data-v-d5abf57a="">(.\*)/月</span>')

findrequirement=re.compile(r'<span class="font" data-v-d5abf57a="">(.\*?)</span>')

* After we were denied to get access to the website, the "User-Agent" was added in the header so the server could recognize that we were not robots and keep scraping.

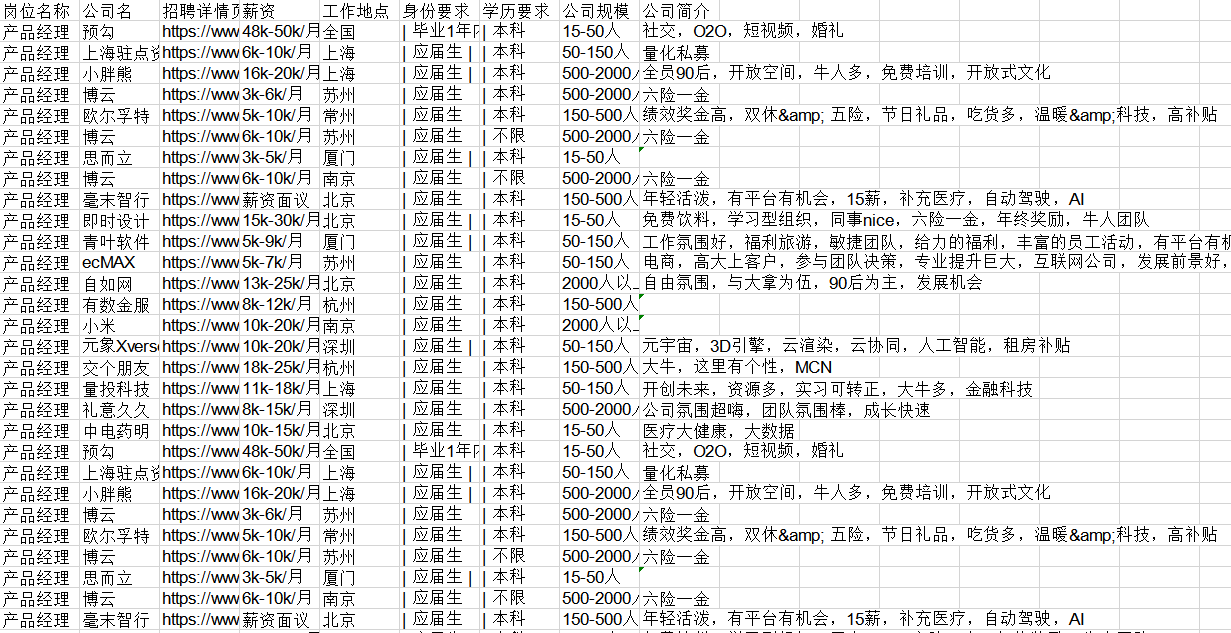
**Figure 4**



***Scraped results***

Parts of scraped result are shown below. Each position contains nine variables "Job Name", "Company Name", "Recruitment Details Page", "Salary", "Work Location", "Identity Requirements", "Education Requirements", "Company Size", "Company Profile", a total of 200 data. We used excel as the format for storing data. In order to facilitate subsequent operations, we manually changed the data from excel to csv.

**Figure 5**



## 3.2 Data cleaning

**Figure 6**

Raw data

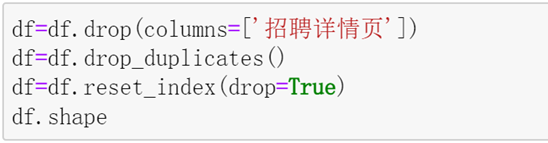
 ***Text processing***

As shown in the raw data in Figure 7, first, we used the pandas drop function to delete the useless "Recruitment Details Page" column. Then, we used the replace function to replace the bar symbols of the two columns of data (string type) of "Identity Requirements" and "Academic Requirements" with spaces. Last. we deleted the duplicate data in the entire raw data frame.

**Figure 7** 

**Figure 8**



**Figure 9** 

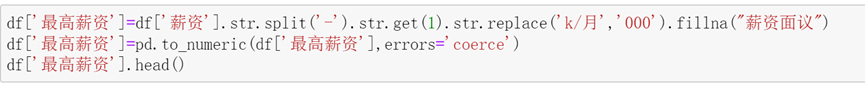
***Numeric values processing***

As shown in Figure 10 and Figure 11, we first divided the string type data in the "salary" column into the "lowest salary" and the "highest salary" using the split function with hyphen notation. Then using the replace function to replace the "k" in the lowest salary column data with "000" and converted its data type to a numeric type. The data in the highest column were processed based on the lowest. If "Salary negotiable" appears in the original data, it will be replaced with a null value. After processing, it will be converted into numeric data.

**Figure 10**



**Figure 11**



## 3.3 Data processing

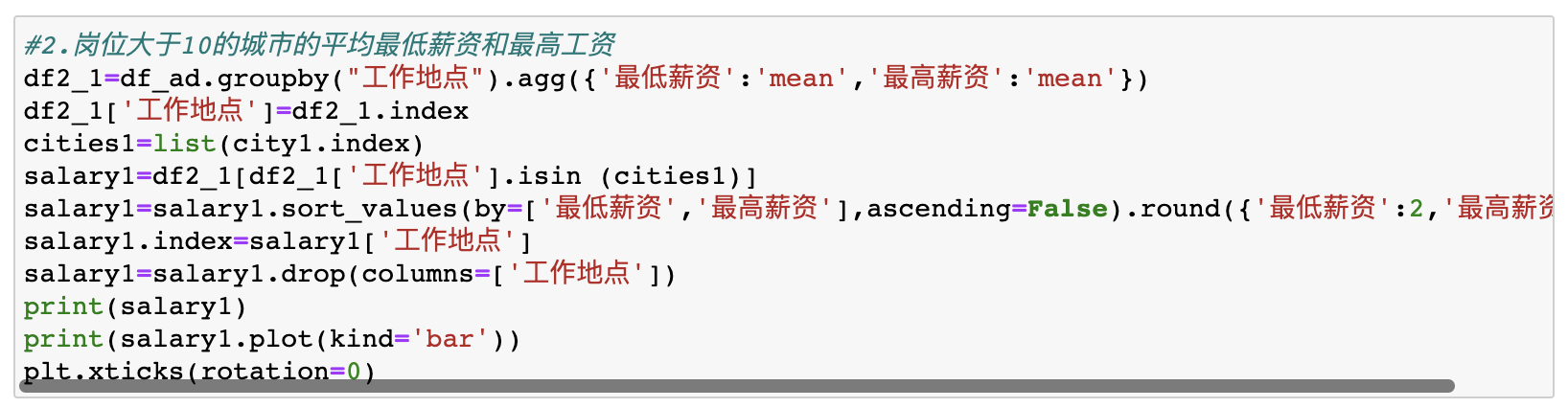
***Visualization processing***

**Figure 12**

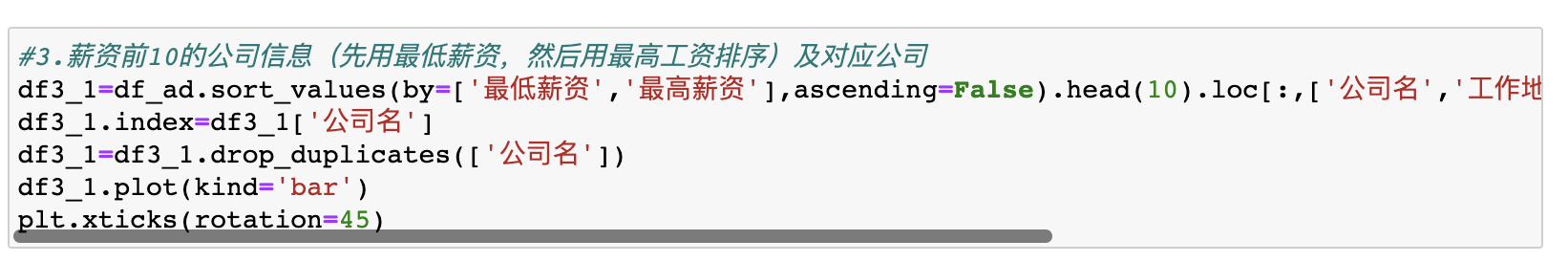


As shown in figure 12, when analyzing ranking of job numbers in each city, first we counted the number of times for each value of the position, which is to see how many positions of each city in total. Then for the number that has been counted, only the top 10 positions were kept, and we extracted statistics which >=10 and save as city 1. After plotting the title, kind, label and rotation of the figure, it will be visualized.

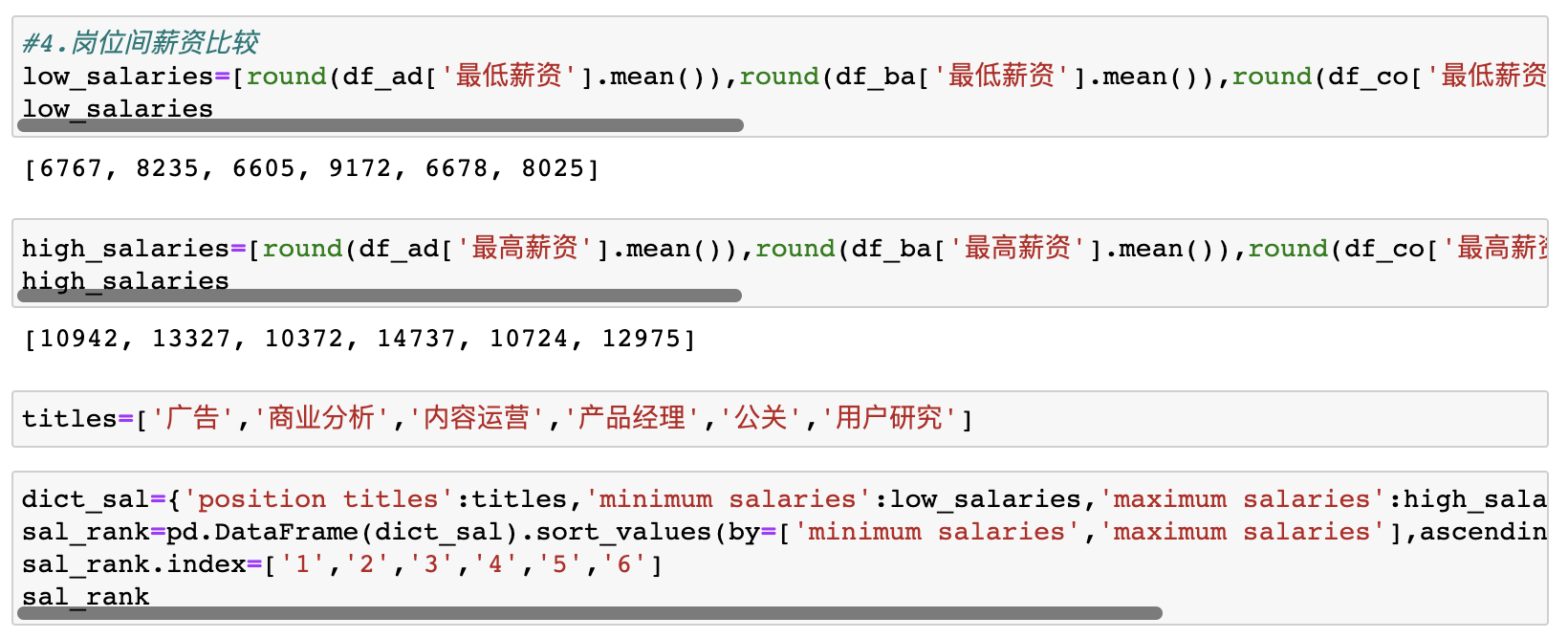
**Figure 13**



**Figure 14**

 For RQ 2 and 3, we first counted the average minimum and maximum salary of different cities and the minimum and maximum salary of different companies. Second, we selected the cities which has over 10 positions and the companies whose salaries are top 10. Finally, we sorted the salary by the minimum one first, and then by the maximum one. The table was plotted according to different cities and companies separately.

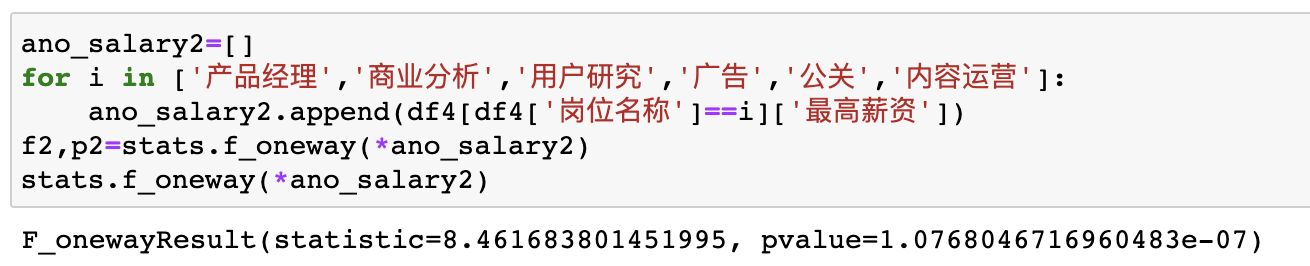
**Figure 15**



**Figure 16**



**Figure 17**



For RQ 4, based on figure 15, first of all, we processed the minimum and maximum salaries for different positions, then named the rows and columns of the chart, and sorted them from high to low according to the minimum and maximum salary.

To check whether there are significant differences among different positions, we did one-way ANOVA analysis twice for the maximum salary and minimum salary of different posts. Based on the figure 16 and 17, the first independent variable is job position, the dependent variable is minimum salaries. The second independent variable is job position, the dependent variable is maximum salaries. The results show that those p-values are all much smaller than 0.05. Hence, the salary difference among different positions are significant.

**Figure 18**

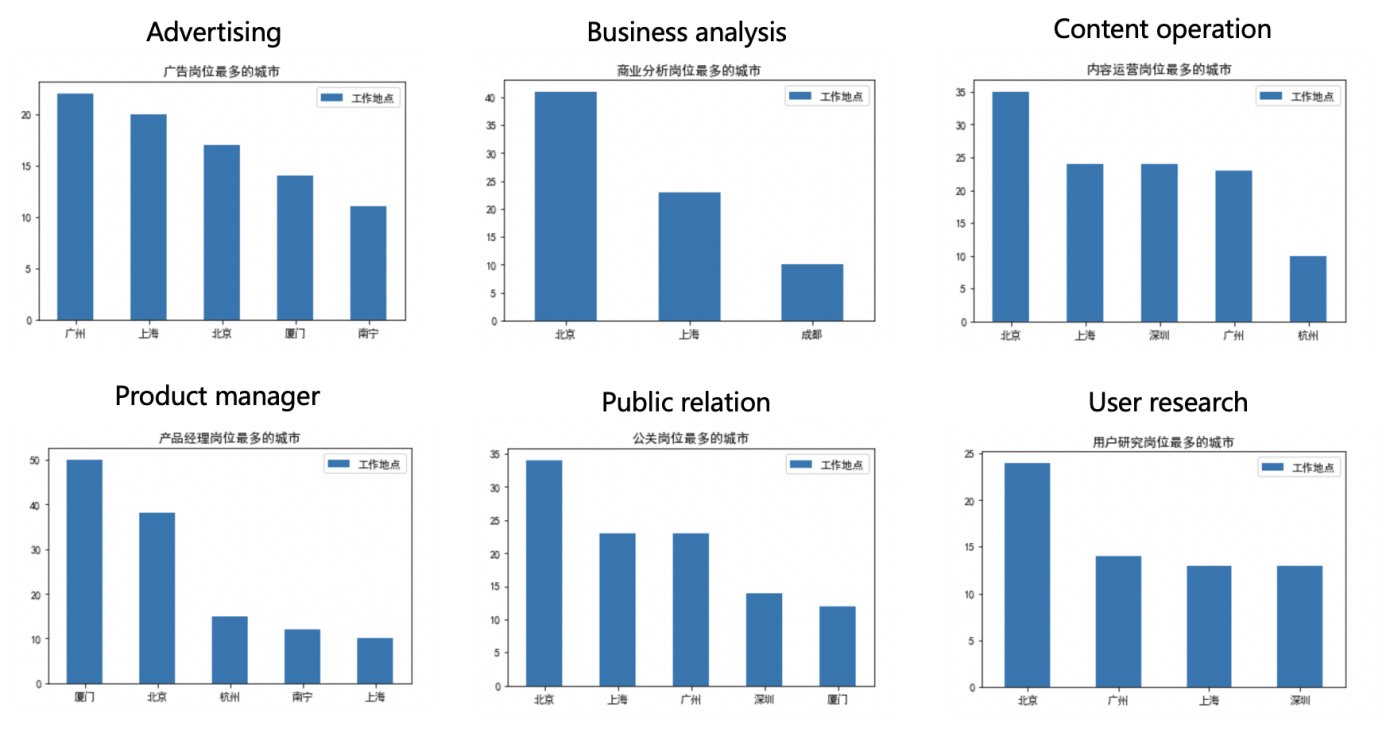


To acquire the word cloud, first, we put background and mask into word cloud. Then, we set the max words and font size and put data into generate. Third, we plotted the word cloud by setting space and visual style and hiding the axis. Finally we used matplotlib.pyplot to show the word cloud.

# 4. Results

**4.1 For each job, which cities provide the largest number(for number>10)?**

**Table 1**

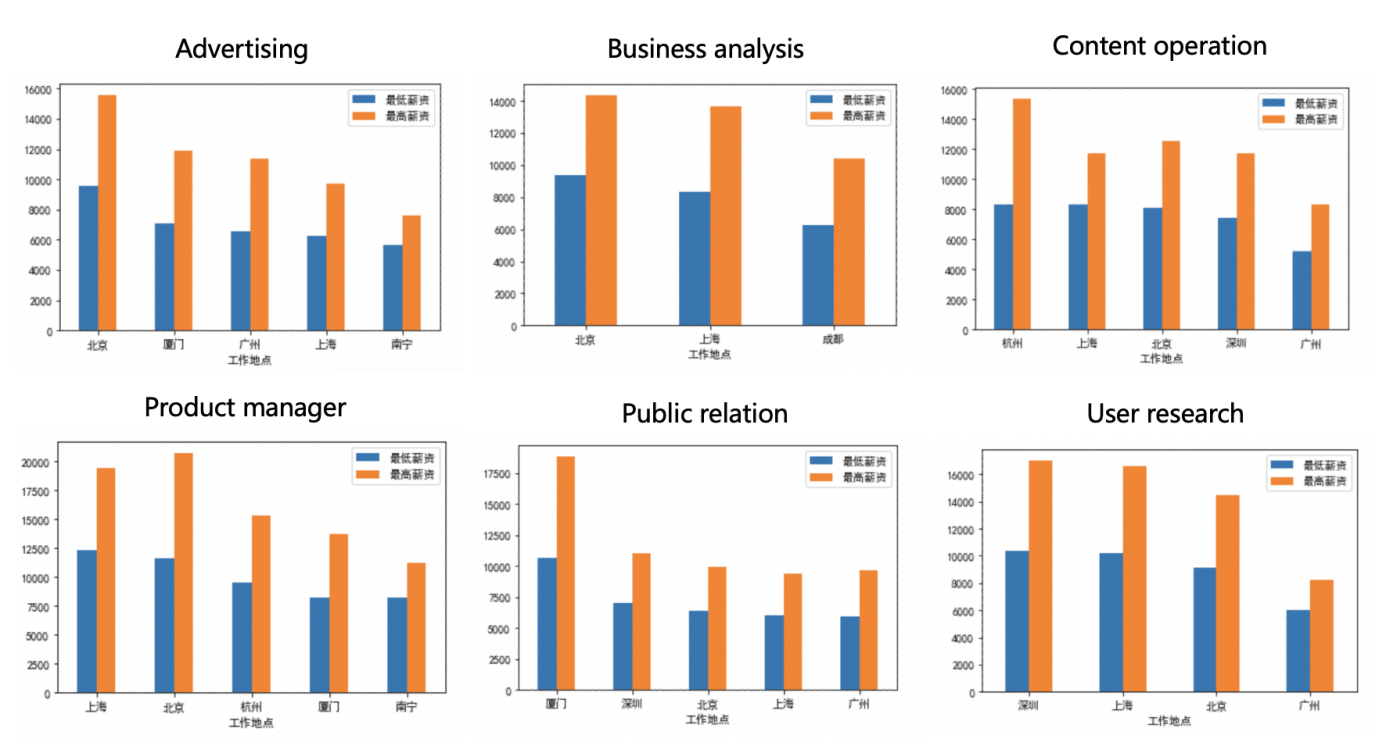


Based on the table 1, for the first question, we concluded that Beijing has the largest job number of business analysis, content operation, public relation and user research, while Xiamen has the largest number of product manager, Guangzhou has the largest number of advertising.

What should be noticed is that only 3 cities including Beijing, Shanghai and Chengdu provide over 10 business analysis positions, and 4 cities including Beijing, Guangzhou, Shanghai and Shenzhen provide over 10 user research positions, which means these two kinds of jobs mostly exist in tier 1 cities and are hard to find in other cities.

**4.2 For cities in RQ1, how much salaries they provide for corresponding job?**

**Table 2**



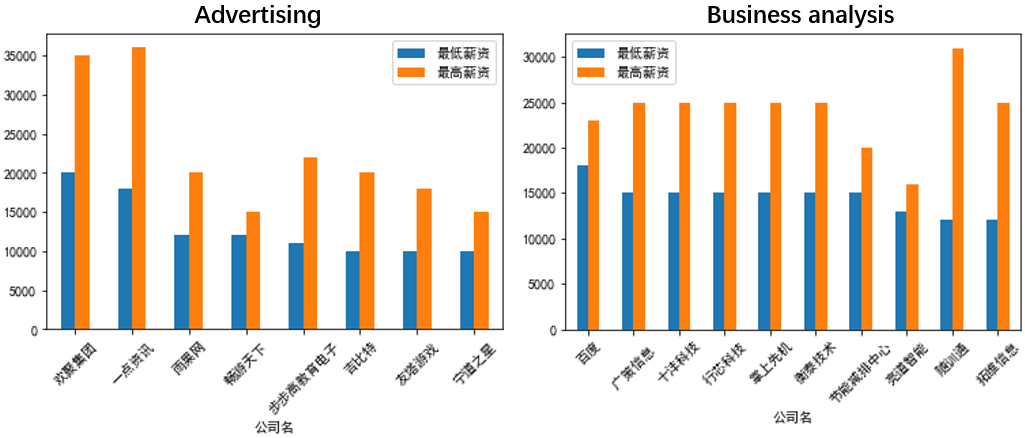
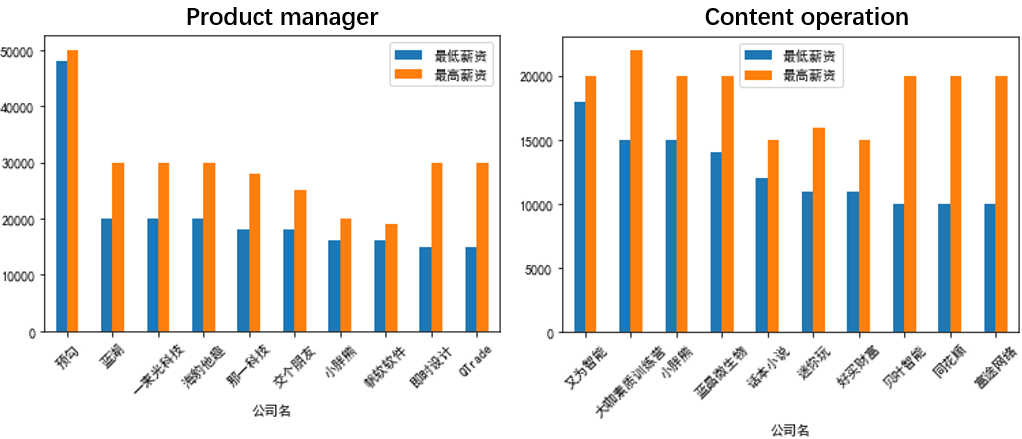
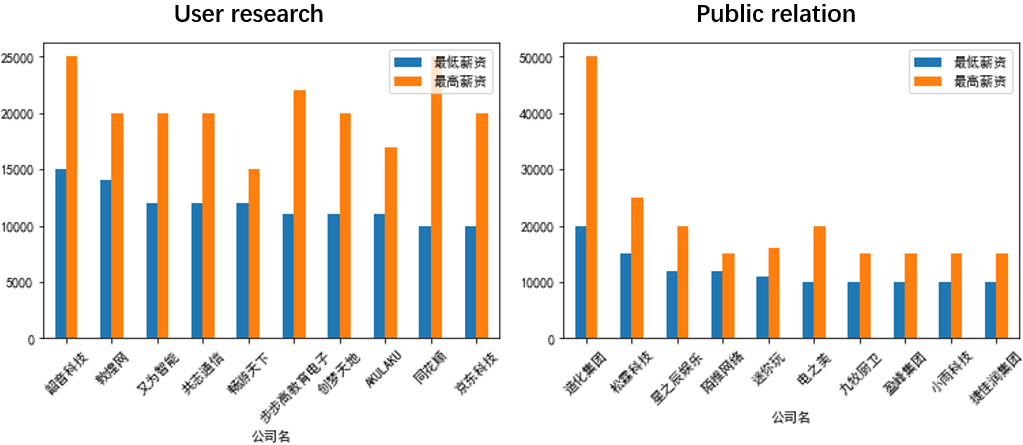
According to Wukong Q&A, there is a question 'Is the salary of 8,000, 10,000, or even fifteen thousand jobs in Zhaopin recruitment real?' Among 348 answers, most respondents say that the real salary they finally get is a little above the minimum salary or even lower. Due to the actual situation, we often considered the minimum salary as the actual one, which implies that the minimum salary is more statistically significant.

Salary was sorted by the minimum salary first, and then by the maximum salary. As shown in the table 2, Hangzhou has the highest minimum salary for content operation. Therefore, if graduates prefer to find a high-salary content operation job, Hangzhou may be a better choice.

Moreover, Beijing has the highest minimum salary for advertising and business analysis. Shanghai has the highest minimum salary for product manager. As for public relation and user research, Xiamen and Shenzhen respectively have the highest minimum salary. In general, first- and second-tier cities can provide higher salaries.

**4.3 For each job, which companies provide the most salaries?​​**

**Table 3**

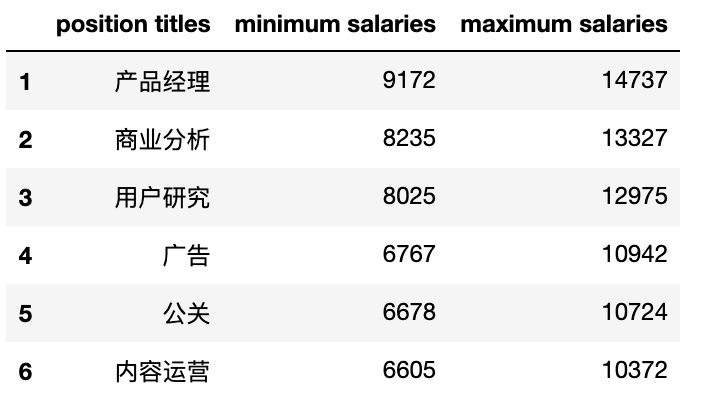
For the third research question, we analyzed the top 10 companies that provide the highest salary for each job.

Based on table 3 above, take product manager as an example, Yugou (预勾) provides the highest minimum salary, which is about 48 thousand. For advertising, Huanju Group (欢聚集团) offers the highest minimum salary, 20 thousand. For business analysis, Baidu (百度) offers the highest salary, which is about 18 thousand. For content operation, Youwei zhineng (又为智能) provides the highest salary, which is around 18 thousand. Lastly, for public relation and user research, Tonghua Group (通化集团) and Shaoyin Techonogy (韶音科技) offer the highest minimum salary respectively.

Those companies shown above may be target companies for whom prefer a higher salary when looking for different jobs.

**4.4 Is there any significant difference in salary among different positions?**

**Table 4**

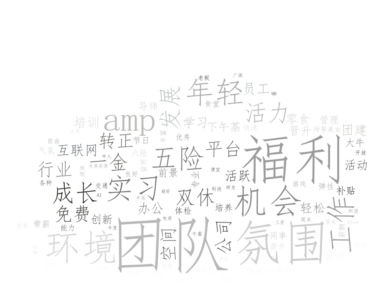


For the last research question, we compared the average maximum and minimum salary of different positions according to the table 4, the highest minimum salary is product manager, which is over 9 thousand. The lowest one is content operation, which is over 6 thousand. The highest maximum salary is product manager too, which is over 14 thousand. The following positions are business analysis, user research, advertising and public relation. Content operation has the lowest maximum salary, but it is still over 10 thousand.

As a result, considering the high salary, product manager is the most worthwhile position for graduates to try.

**4.5 Word cloud**

**Figure 19**



Finally, we analyzed the word frequency of the company introduction and created the word cloud. The five most frequently occurring key words are team, welfare, atmosphere, environment, and chance, which indicates that the graduates focus more on these five factors when finding job, and companies prefer to use these words to attract graduates.

# 5. Conclusions

**5.1 Discussion of findings**

In order to explore the different salary of each position, the salary of each job in different cities and the salary of each job in different companies on ‘Shixiseng’, this paper used Python to crawl, clean and visualize the data. Through crawling the data by four packages including BeautifulSoup, re, requeat and xlwt, this research tends to study on how to help the graduates to find a satisfying job, and then this paper has utilized strip, replace, plot and word cloud to do data cleaning and data processing.

According to the result, we found that there is a significant difference in salary between different positions for the previous graduates of the Department of Media and Communication at City University of Hong Kong.

First of all, it is obvious that tier 1 cities such as Beijing, Guangzhou, and Shanghai provide more job opportunities. This is one of the reasons why students tend to go to big cities after graduation. However, positions with high professional requirements such as business analysis even only exist in Beijing, Shanghai and Chengdu.

Secondly, for different positions, the salary gap in each city is large. Generally speaking, the more developed the city is, the higher the salary they may provide. But there are also differences in different positions. Then, if students tend to find high-paying jobs, they can also focus on the top ten companies that we have screened out for different positions.

Finally, through analysis, it is found that the salary of product managers is generally higher than that of other positions, which may provide a reasonable employment direction for students. Moreover, team, welfare and atmosphere are top 3 factors which influence the employment choice of graduates. It shows that graduates pay more attention on the style the companies have.

As a result, the founding we discovered may assist the new media graduates in both City University of Hong Kong and other universities who prefer to find a job in mainland China, allowing them to gain a certain reference for choosing work, working cities and companies.

**5.2 Limitations and improvements**

***The platform can be more various***

Providing college students with more than 250,000 corporate internships, 'Shixiseng' is selected as our research data platform, but not all graduates will use this job search platform, which makes the data may not be of reference. At the same time, the research data does not apply to students who tend to seek jobs in Hong Kong. Therefore, if the data collection platform is extended to Zhilian, Liepin, LinkedIn, etc., the accuracy of the research will be improved to a certain extent, and this will also provide graduates with better references.

In addition, only 200 pieces of data for each position may not be convincing enough, but the number of positions for “Shixiseng” is limited, and the data after 200 pieces are often inaccurate, which will reduce credibility. Therefore, if it can be verified with the information of other domestic recruitment websites, such as “BOSS Zhipin”, the accuracy of the conclusion will be improved.

***Effectively induct and exclude false data***

When analyzing the third research question, we found that the Yugou company’s minimum salary for product manager positions, which is 48 thousand, is far higher than other companies. Therefore, we inferred that this salary may be an invalid sample and a false figure used by the company to attract job applicants. It is necessary for us to filter the false data based on the actual situation in order not to influence the results.

**Reference**

Xinhua Net (Ed.) (2021). *Graduate Employability Research Report 2021.*

<http://www.xinhuanet.com/tech/2021-06/11/c_1127555241.htm>

Yicai (2021). *Average income in 40 big cities: Seven cities exceed sixteen thousand. Yangtze River Delta is richest.* https://www.yicai.com/news/101020119