Report on Data Analyst Recruitment Dataset

Statistical Computation and Software Fall 2021

Group 15: Mao Wenhui, Xu Tianyi, Jian Xinyao, Yang Letian

Abstract

The National 14th Five-Year Plan proposes to accelerate digital development, build a digital China, and urgently recruit digital talents across the industry. Many students in the Department of Statistics hope to devote themselves to the field of data analysis in the future. We believe that only by clearly understanding the current needs of enterprises for data analysis positions can we ensure that the direction we learn is in line with the actual needs of the enterprise.

This report is based on 10,266 recruitment information crawled on a certain large-scale recruitment website with "data analysis" and "business analysis" as keywords, and uses R language programming for comprehensive analysis. This project studies the effect of education on finding a job; explores cities that are more conducive to the development of data analysts; and discusses the differences in wage, welfare and requirements between companies of different sizes. The above issues were explored in detail by visualizing the data (bar charts, box plots, pie charts) and performing statistical significance tests (Welch F test, Game Howell test) and also using regression prediction (logistic regression, random forest, neural networks).

The conclusions are as followed: Bachelor degree can meet the requirements of more than 95% of enterprises, experience is more important than education; cities in Tier 1 provide higher wages and more jobs and Beijing is the best country for data analysts; large companies provide higher wages and more welfare but their demands are lower than small companies and have stricter requirement for education and work experience; the importance of variables is experience, city level, education, and company size.

Keywords: Data analyst recruitment, Welch's F test, predictions

Contents

1 Introduction

According to the “2020 China Big Data Industry Development White Paper” issued by China Academy of Information and Communications Technology(CAICT, 2020), the scale of China’s big data industry is expected to reach 1,016.66 billion Yuan in 2022. From 2017 to 2022, the 5-year compound growth rate reached 23.3%. At the same time, with the increasing demand for data analysis jobs, the supply of core big data talents in China has been far lower than market demand.

Facing the massive amount of information released on the recruitment website, it is difficult for job seekers to see the general trend of job demand, thus not knowing how they should improve themselves and make the right choice.

Therefore, this report is committed to doing a comprehensive and in-depth analysis of the current situation of data analysis positions in the recruitment website and some important influencing factors. The results can help relevant job seekers to see the current situation of data analysis recruitment and the future trend.

2 Research Objectives

Explore the influence factors of data analysts’ mean wage in three perspectives.

- Study the effect of education on finding a data analyst job.
- Explore cities that are more conducive to the development of data analysts.
- Study the differences in wage, welfare and requirements between companies of different sizes.

3 Data Processing

3.1 Data Preprocessing

The data analysis job recruitment data set “jobinfo.csv” consists 10,266 data and 16 variables. There are a small amount of duplicate, invalid and missing information. The data needs to be preprocessed for further study. The exact adjustments are as follows:

- Delete the duplicate data.
- Since more than 20 cities are hard for analyze, we combined the given data “city_level.csv” with “jobinfo.csv” to get the corresponding city level of each entry, adding the new variable ”level”.

- Next, we fill in missing data. We assign the NA and "若干" in "num_people" of 1 and 3 respectively. For the NA in experience, we assign it of 0, meaning no requirement for experience. For NA in education, we assign it of "无要求".
- Delete the entries with missing company size and company type.
- Use function "boxplot()$out" to delete the outliers of mean wage.

3.2 Text Analysis

Add two new variables "skill_count" and "wel_count" by following method:

Use the function *segment()* and the provided "stopwords.txt" and "add_dict.txt" in the *jiebaR* package to segment the "welfare" and "job_intro" columns. For the "job_intro" column, the main purpose is to extract the English words in the text, which are usually related to a certain skill (e.g. SPSS, SQL). For the "welfare" column, since it is separated by vertical lines, the *strsplit()* function can be used to complete the word splitting. After the word splitting, word frequency tables were made and the top 20 keywords of skills and welfare were selected. The two columns added are the occurrence frequency of TOP 20 of skills and welfare keywords, respectively.(e.g. some entry only contains "SQL" in the job_intro column, then its skill_count is 1) The specific index names and descriptions are shown in Table 1 below(Joffy Z., 2018):

Indicator	**Type**	**Description**
level	categorical	The city level the company bases on
industry_1	categorical	The industry of the company E.g. Computer/Internet
experience	numerical	The requirement of working experience [0:10]
num_people	numerical	Number of recruits [1:280]
education	categorical	The requirement of education level
company_type	categorical	The type of the company E.g.private company
company_size	categorical	Number of employees in the company
mean_wage	numerical	$wage_{mean} = \frac{wage_{max}+wage_{min}}{2}$
skill_count	numerical	Number of skill keywords mentioned in "job_intro"
wel_count	numerical	Number of welfare keywords mention in "welfare"

Table 1: Factors and description

4 Major Findings

4.1 Impact of Education in Job Recruitment

- **Background**

As we know, job recruitment often has certain requirements for education. Candidates with different educational backgrounds may have differences in demand of recruitment, job skills requirements, and wages. According to these differences, we try to study the impact of education in recruitment information based on this data set, the following are our analysis.

4.1.1 Overview of the Data

- **Distribution of Education in Demand**

Education	Freq
中技	9
中专	118
高中	50
无要求	1066
大专	2370
本科	5054
硕士	386
博士	6

Table 2: Education with frequency

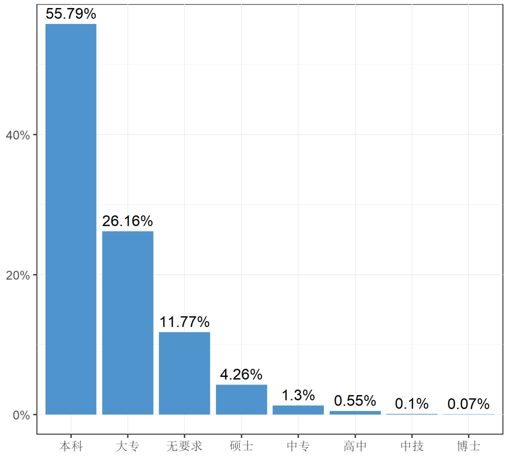

Table 3: Education with the percentage

From the table 2 and table 3, we can know that the demand for the bachelor's degree(“本科”) is the largest, reaching 55.79%, followed by the technical college degree(“大专”), accounting for 26.16%. They are also 11.77% missing values(“无要求”), ranked third. The need for the master's degree(“硕士”) is relatively small, only accounting for 4.26%. Then the demand for the technical secondary degree(“中专”), high school degree(“高中”), and vocational secondary degree(“中技”) decrease accordingly, they account for about 2% in total. The recruitment in this data required for a doctor's degree(“博士”) is less than 0.1%.

Therefore, based on this result, we can know that we can meet more than 95% of the educational requirements in recruitment when we have a bachelor's degree.

- **Distribution of Education in City Level**

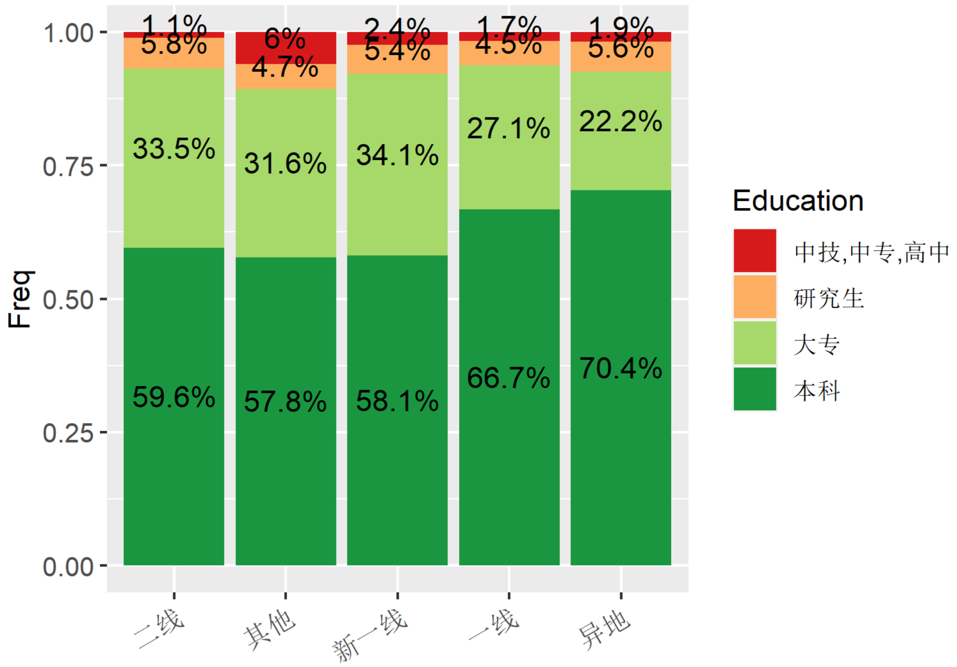

Figure 1: Education in different city levels

Will different cities have different requirements for education?

Because the sample sizes of the vocational secondary degree, technical secondary degree, high school degree, doctor's degree are relatively small, we combine vocational secondary degree, technical secondary degree, high school degree into one category, and combine the master's degree and doctor's degree into the graduate degree.

From the figure 1, we can see that the Tier 1, Tier2 cities of different levels have roughly the same demand for different educations, which are mainly undergraduates. Relatively speaking, the Tier 1 cities have a higher demand for undergraduates, and the Tier 3, Tier 4, and other tiers cities have a higher acceptance rate of "vocational secondary degree, technical secondary degree, high school degree" with 6%.

- **Distribution of wage with education**

From the table 4 and box plot 5, we can see that the wage has also increased with the increase in education. From a preliminary point of view, according to the mean_wage from low to high, the corresponding degree of education is the vocational secondary degree, technical college degree, high school degree, technical secondary degree, bachelor's degree, master's degree, doctor's degree.

Obviously, the wage of doctor's degree is much higher than others, the wage of bachelor's and master's are similar, the wage of technical college's and high school's are similar.

- **Significance of Education on Wage**

Education	min_wage	mean_wage	max_wage
中技	4711.11	6772.22	8833.33
中专	7484.46	9866.94	12249.43
高中	6110.00	8030.00	9950.00
大专	6422.37	7999.06	9575.77
本科	8912.50	11506.56	14100.64
硕士	9490.88	12211.80	14932.74
博士	13333.33	17500.00	21777.67

Table 4: The table of wages with education.

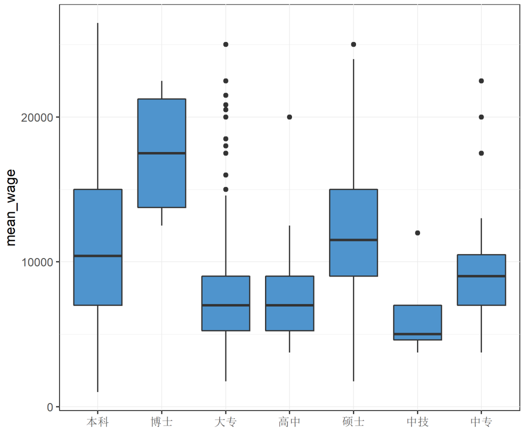

Table 5: Boxplot of mean_wage with education

Next, we want to test whether the impact of education on wages is significant.

Because the sample sizes of the vocational secondary degree, technical secondary degree, high school degree, doctor's degree are relatively small, we combine vocational secondary degree, technical secondary degree, high school degree into one category, and combine the master's degree and doctor's degree into the graduate degree.

Since one-way ANOVA has two assumptions, normality and homoscedasticity, we try qqPlot and bartlett.test, then find it can not meet the homoscedasticity. Then we call the welch F test in onewaytests. The result in table 6 shows p-value=2.875034e-135<0.05, which indicates the difference of mean wage between different education is statistically significant.

data	mean_wage and education
statistic :	356.4749
num df :	3
denom df :	629.8788
p.value :	2.875034e-135
Result :	Difference is statistically significant.

Table 6: Welch's heteroscedastic F test (alpha = 0.05)

The games Howell test is conducted to figure out exactly which groups have the difference, which is an improved version of the Tukey-Kramer method and is applicable in cases where the equivalence of variance assumption is violated; it is a t-test using Welch's degree of freedom(Lee, S. Lee, D. K., 2018).

From the table 7, all groups with different education have a significant difference in mean wage.

.y.	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
mean_wage	本科	大专	-3507.50	-3794.77	-3220.23	2.48e-13	****
mean_wage	本科	研究生	786.18	79.70	1492.66	2.20e-02	*
mean_wage	本科	中技，中专，高中	-2315.88	-3232.39	-1399.39	3.05e-09	****
mean_wage	大专	研究生	4293.67	3583.94	5003.41	1.89e-11	****
mean_wage	大专	中技，中专，高中	1191.61	272.62	2110.59	5.00e-03	**
mean_wage	研究生	中技，中专，高中	-3102.06	-4221.95	-1982.18	1.49e-11	****

Table 7: games_howell_test

4.1.2 Comparison with Experience

Next, we want to know the difference between the impact of education and experience on salary. Because of the heteroscedasticity between groups, we call gpTwoWay in twowaytests (to solve the zero items in the classification and the singularity in the calculation, the experience value is 3,5,8 combined). The result in figure 2 is education, experience, and cross-terms all have a significant impact on wages.

Factor	P-value	Result
experience	0	Reject
education	0	Reject
experience:education	0	Reject

Figure 2: Result of PB method(α= 0.05)

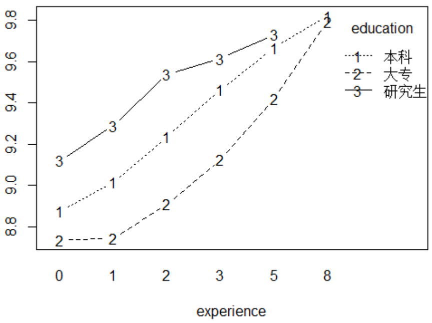

Figure 3: Interaction plot of experience and education

Through interaction.plot figure 3, we can also see that when the number of years of experience required increases, the increase rate of mean wage for different education is different.

To better compare the impact of experience and education on wage, we convert the education to the corresponding number of years, then turn it into a continuous variable. Here, to better define this variable, we only extract the data of bachelor's degree, master's degree, and doctor's degree, then define them as 4,6 and 10 in turn.

In the regression results in table 8, the coefficient of experience is 0.168, the coefficient of education is

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.453992	0.046525	181.71	<2e-16	***
education	0.110152	0.010737	10.26	<2e-16	***
experience	0.168080	0.003793	44.31	<2e-16	***

Table 8: Linear regression outcome

0.11, which is smaller than experience's. Under this data model, the impact of experience on wages is more significant.

4.1.3 Requirement of Skills

Finally, we explore the difference between the requirements for skills of different education in the data.

Distinguished by bachelor's degree, we count the Top 10 required skills under vocational secondary, technical secondary, high school, and technical college degree and TOP10 required skills under bachelor's, master's, and doctor's degree.

x	freq	percent
EXCEL	1082	42.48%
SQL	434	17.04%
PPT	292	11.46%
OFFICE	280	10.99%
PYTHON	255	10.01%
SPSS	188	7.38%
BI	187	7.34%
SAS	167	6.56%
WORD	167	6.56%
R	143	5.61%

Table 9: Top 10 skills of under bachelor degree

x	freq	percent
SQL	2139	39.28%
PYTHON	1772	32.54%
EXCEL	1768	32.46%
R	1170	21.48%
SPSS	808	14.84%
SAS	746	13.7%
PPT	737	13.53%
BI	719	13.2%
HIVE	469	8.61%
OFFICE	439	8.06%

Table 10: Top 10 skills of above bachelor degree

As can be seen from the table 9 on the left, 42.48% of the recruitment requires the mastery of excel, 17.04% of the recruitment requires the ability of the SQL, followed by PPT, OFFICE, PYTHON.The following are minor than 10%, while there is a big difference in the table of skills required with a bachelor or above degree.

The SQL ranks first in the right table 10, with nearly 40%, followed by PYTHON, EXCEL, R, SPSS, SAS, etc., which all account for more than 10%. So it can be seen from this comparison that the bachelor's degree or above requires relatively more skills.

4.1.4 Education with Prediction

As an undergraduate, we care about which jobs require a bachelor's degree or above. Hence we hope to build a classification model to help us identify.

We use logistic regression and decision tree model respectively, and the predictions are as follows. In this data set, the accuracy of logistic regression (73.6%) is higher than that of decision trees (71.7%)

Act-Pred	本科及以上	本科以下
本科及以上	1429	209
本科以下	424	336

Table 11: Confusion matrix of logit model

Act-Pred	本科及以上	本科以下
本科及以上	1405	233
本科以下	445	315

Table 12: Confusion matrix of decision tree

4.1.5 Summary

Through the above analysis of education, we know that there are many employment opportunities for a job seeker with a bachelor's degree or above, but there are correspondingly more skills requirements. The proportion of demand for talents in different cities is roughly the same. Accumulating experience will significantly help increase wages.

4.2 Differences in Data Analysis Positions Based on Geography Area

- **Background**

In China, the overall level of employment, as well as the overall wage may vary from city to city. Here, we only explore the differences in data analyst and business analyst jobs in different cities in China and the corresponding average wage differences.

The data analyst position is still evolving as a newly developing industry. At the same time, it has interactions with various industries. Here, we explore the differences in demand and mean wage for data analyst-related jobs in each region and draw conclusions to provide advice on location selection for those interested in future employment planning.

Level	Freq	Percent
Tier 1 cities（一线）	5531	56.64%
New Tier 1 cities （新一线）	2965	30.36%
Tier 2 cities (二线)	1038	10.63%
Tier 3 cities （三线）	158	1.62%
Tier 4 cities （四线）	49	0.50%
Tier 5 cities （五线）	25	0.26%

Table 13: City level before merging

Note: Differences in consumption levels and the average cost of living between cities are not considered here.

4.2.1 Overview of the Data

- **City**

In the file “jobinfo.csv”, ‘city’ is a ‘char’ variable with no missing values, and there are 143 categories in total. Among them, regional units are not uniform, including provincial (e.g., Zhejiang Province), municipal (e.g., Hangzhou), and off-site recruitment. Considering its many categories, and the number of samples under many categories is too small(only one or two, accounting for only 0.01103874% of the total), by calculating the ratio of sample amounts of each city to the total sample amounts, we merge the cities with a ratio of less than 0.005 into the “other” category. After the merging, there are 27 categories left, including “off-site recruitment” and “other”.

- **City Level**

In the file “city_level.csv”, ‘level’ is a ‘char’ variable, and the “2019 City Business Charm Ranking” released by First Finance in 2019 divides all prefecture-level cities in China into six levels, including Tier 1, new Tier 1, Tier 2, Tier 3, Tier 4, and Tier 5, as shown in the table 13. Base on the merged result above(with 27 categories), the cities were further merged into five categories: Tier 1, new Tier 1, Tier 2, Offsite Recruitment, and Other.

4.2.2 Levels’ Difference : Mean Wage and Demand

First, the chart gives a general sense of the differences between the different city levels.

- **Demand**

The figure 4 shows the difference in demand between different levels. Among them, Tier 1 cities (light

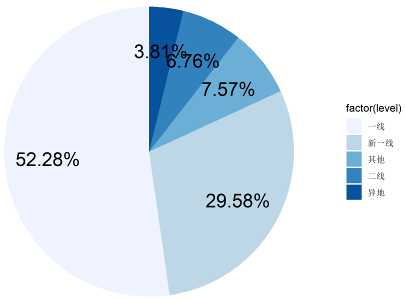

Figure 4: Distribution of city level

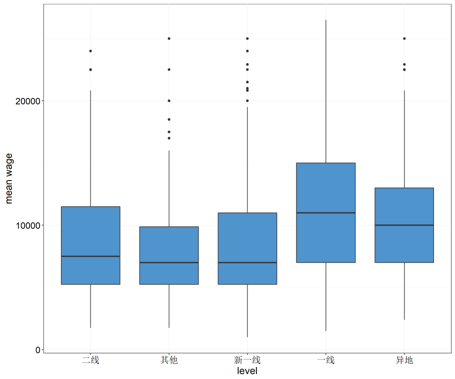

Figure 5: Box plot of mean wage v.s. city level

blue) account for the highest proportion, accounting for 52.8% of all demand; new Tier 1 cities are second, accounting for 29.58%, Tier 2 and other levels of cities account for similar proportions, while the demand for off-site recruitment is the least, accounting for only 3.81% of the total.

- **Mean Wage**

The figure 5 shows the distribution of mean wages in each city level. In general, the mean wage in Tier 1 cities is generally higher, followed by off-site recruitment, Tier 2 cities, and new Tier 1 cities are at a similar level, and the overall wage of data analyst positions in other cities is low.

With the one-way ANOVA, we can verify whether there is a significant difference in the average wage between different city levels. Before that, we test the two prerequisite assumptions of ANOVA, namely the dependent variable is assumed to be normally distributed, and has equal variance in each group.

When testing the homogeneity of variances, the p-value of *bartlett.test* is 1.339685e-164 which is close to zero, thus the assumption of equal variance is not satisfied. So instead we use Welch's Heteroscedastic F test and the result is statistically significant($p < 0.0001$), providing strong evidence that data analyst positions do vary by levels of the city. And multiple comparison procedure shows how specific city levels differ from each other. *Games_howell_test()* indicates that there is significant difference between levels except for the new Tier 1 cities and Tier 2 cities, which is also consistent with the graph 5.

Combining the analysis of wage and demand, the following conclusions can be drawn at this stage at the level of the urban hierarchy.

1. Tier 1 cities have the most demand, and generally higher wages, are the preferred region for job hunting.
2. New Tier 1 cities are in second place in terms of demand, but generally have lower wages.
3. There is less demand for off-site recruitment, but higher wages.

4.2.3 Cities' Difference : Mean Wage and Demand

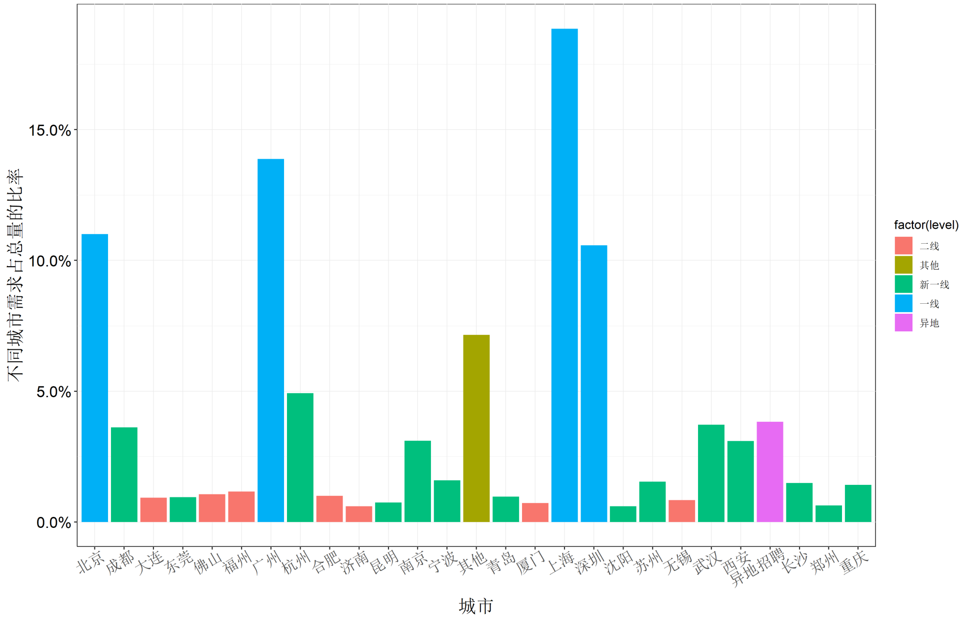

Figure 6: Demand of cities

- **Demand**

From demand perspective, the figure 6 reveals the following findings, which are largely consistent with the findings at the city level.

1. The Tier 1 cities (blue) each account for 10% of the demand, with Shanghai's share even reaching nearly 20%.

2. Except for Hangzhou, there is no particularly significant difference in demand between generally new Tier 1 cities (green) and Tier 2 cities (red).

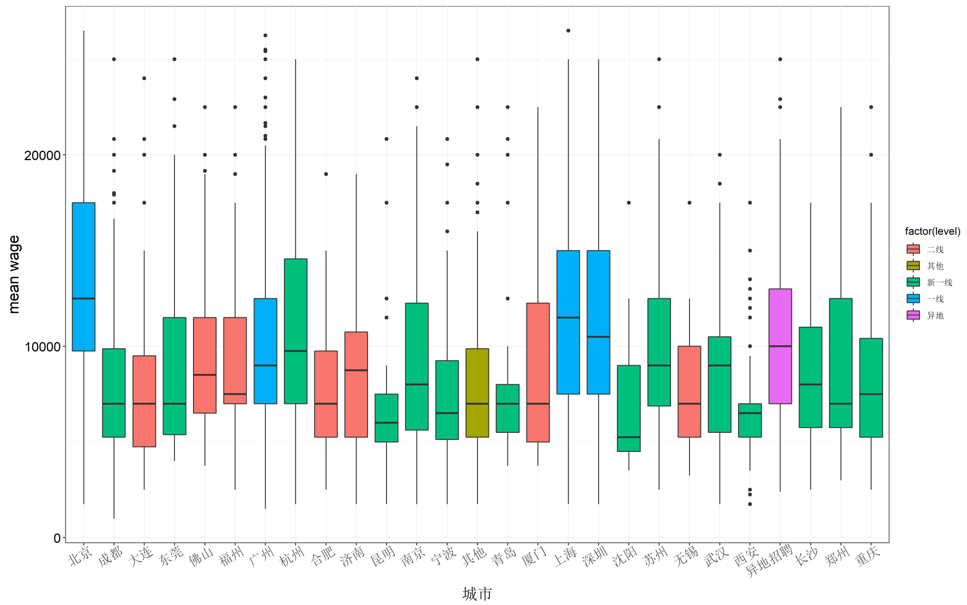

Figure 7: Mean wage of cities

- **Mean Wage**

From the perspective of the mean wage, the following conclusions can be drawn from the figure 7.

1. The mean wages in Tier 1 cities (blue) are significantly higher than other cities, basically the mean monthly wage is more than 10,000, except for Guangzhou, whose mean wage is similar to the Tier 2 city of Xiamen, the new Tier 1 cities of Nanjing, Suzhou, and Zhengzhou.

2. Among the new Tier 1 cities (green), except Hangzhou has an overall higher wage, the others are basically similar to Tier 2 cities.

3. The overall wages for off-site recruitment are also high, at around 10,000.

4. Shenyang, a new Tier 1 city, however, has the lowest mean wage. Other cities with lower wages are Kunming and Xi'an, which indicate that the mean wage of data analysis posts may be related to the geographical location of the city.

Combining wage's variance and span gives a more comprehensive consideration. If only the span is considered it may lead to some biased conclusions. For example, if we only consider the wage span of a certain city, we can find that though the mean wages of Tier 1 cities are high, their spans are also large, and the minimum of wage is equal to or even lower than which of other cities. In the new Tier 1 cities Dongguan, Qingdao, and Tier 2 city Xiamen, their minimum mean wages are relatively high, from this point of view, work in Dongguan, Qingdao and Xiamen, wage will be more secure. If we consider the variance together, the following figure 8 shows that in Beijing and Xiamen, for example, although the wage span is larger in Beijing, it has a small confidence interval, which means that the mean wage in Beijing is more concentrated and the probability of getting the lowest mean wage is very low, while in Xiamen the wage is relatively dispersed and the probability of getting a low wage is higher. The confidence interval of wage distribution in Beijing is higher than that in Xiamen, which means that the conditions in Beijing are more favorable than those in Xiamen in terms of wage.

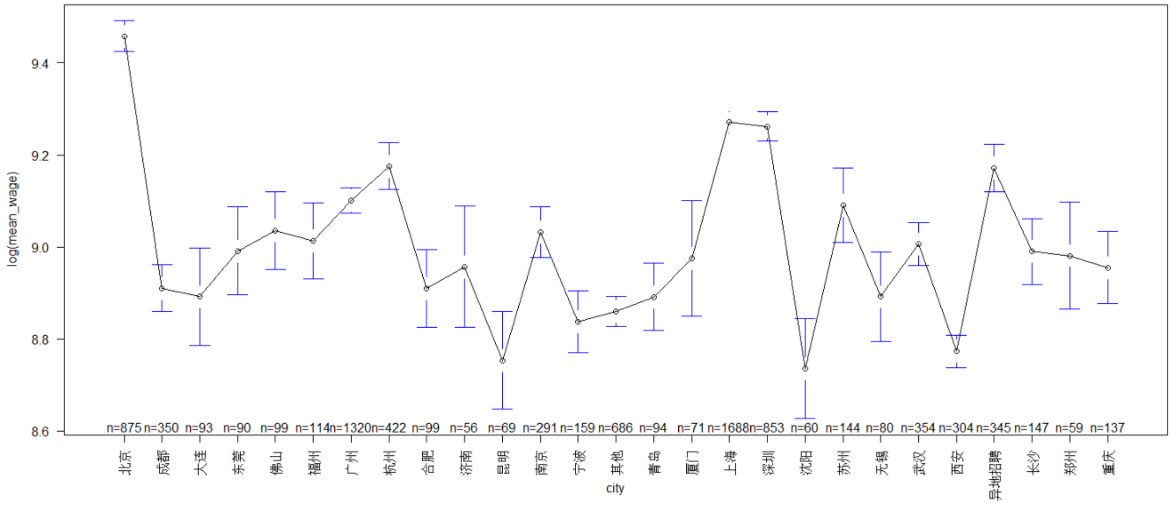

Figure 8: Mean and CI of cities' wage

Combining the above factors, the Tier 1 cities (Beijing, Shanghai, Guangzhou, Shenzhen) and Hangzhou, which is a new Tier 1 city, can be recognized as a more desirable place to seek employment.

The premise hypothesis of the ANOVA for wages in these five cities was found to be satisfied, and the ANOVA F test shows that there are differences between wages in these five cities. Considering the differences in the sample sizes of the cities, a two-by-two comparison of the cities using *Scheffe's test* reveals that there is no significant difference between Guangzhou and Hangzhou, no significant difference between Shenzhen and Shanghai, and no particularly significant difference between Shanghai, Shenzhen, and Hangzhou. Accordingly, the following recommendations are given.

1. Data analytics positions in Beijing are extremely well-paid and in high demand.

2. Data analytics jobs in Shanghai are second to Beijing only in terms of salary, and Shanghai has the highest demand for such jobs compared to other cities, which makes it no less of an option than Beijing on balance.

3. The mean wages in Shenzhen and Shanghai are similar, with Shanghai slightly higher. Meanwhile,

Shanghai has nearly twice the demand for Shenzhen and a smaller confidence interval, so the choice of job search in Shanghai than Shenzhen should be more priority.

4. The mean wage in Guangzhou and Hangzhou are similar, with Hangzhou being slightly higher. However, Guangzhou has nearly three times the demand of Hangzhou, which means there are more opportunities and choices in Guangzhou.

4.2.4 Cities, Industries and Skills

Different cities will have different biased industries, for example, the northeast is biased towards heavy industry and the coast is biased towards the light industry. As can be seen from the figure 9, for data analysis positions, generally, the computer/Internet/communication/electronics(计算机/互联网/通信/电子) industry has the highest demand, followed by trade/consumer/manufacturing/operating(贸易/消费/制造/营运) industry. For professional services/education/training (专业服务/教育/培训) and accounting/finance/banking/insurance(会计/金融/银行/保险), etc, there is also a small demand. Other industries have less demand. The difference between cities is that in Hangzhou, the demand for computer/Internet/communication/electronics industry and trade/consumer/manufacturing/operating industry is more similar; while in Beijing, accounting/finance/banking/insurance industry has more demand compared to trade/consumer/manufacturing/operating.

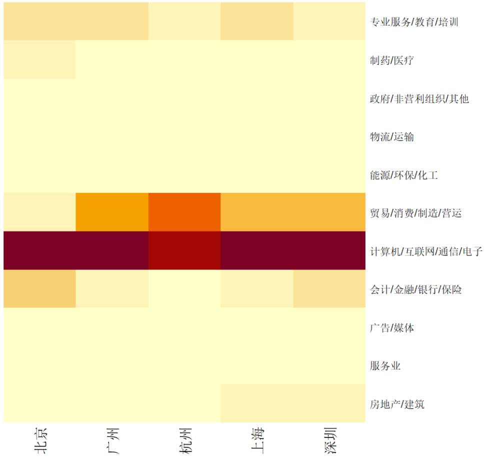

Figure 9: Heatmap : cities and industries

Generally speaking, the skills corresponding to different industries will also vary. Here, take the computer/internet/communication/electronics industry as an example, through statistical word frequency, table 14 shows that for data analysis positions, the main skills required by computer/internet/communication/electronics and other industries are similar, mainly including SQL, EXCEL, PYTHON, R. The specific differences between industries are that the computer/internet/communication/electronics industry may require more hive, Hadoop, SPARK, Oracle, ETL, such large-scale data processing tools, while others will focus on the requirements of management (such as CRM), marketing skills.

	main	others
1	SQL	EXCEL
2	EXCEL	SQL
3	PYTHON	PYTHON
4	RSTUDIO	DATA
5	DBA	RSTUDIO
6	SPSS	PPT
7	BI	BI
8	PPT	DBA
9	SAS	SAS
10	HIVE	SPSS

Table 14: Top 10 skills

	main	other
1	HIVE	BUSINESS
2	HADOOP	EXPERIENCE
3	SPARK	WORD
4	JAVA	ANALYSIS
5	ORACLE	CRM
6	ETL	MARKETING
7	LINUX	VBA

Table 15: Unique skills

4.2.5 Summary

Overall, the Tier 1 cities from the mean wage and demand for consideration are better job search areas, including Beijing, Shanghai conditions are optimal. In the five cities of Beijing, Shanghai, Guangzhou, Shenzhen, and Hangzhou, computer / Internet/communications/electronics industry demand for data analysis positions are the largest, the skills required in addition to the general requirements of SQL, EXCEL, PYTHON, R, but also included some large-scale data processing tools.

4.3 Differences in data analysis positions based on company size

Nowadays, it is very popular to find a job in a “big company”. The so-called “big company” refers to companies that are well-known, large in scale, and have histories. People often think that working in a big company ensure a better salary. But at the same time, people also need to consider the high standards and competitive pressures of these companies when finding a job.

4.3.1 Overview of the Data

There are 7 types of company sizes:A. less than 50 people, B.50-150, C.150-500, D.500-1000, E.1000-5000, F.5000-10000, G.more than 10000 people.

4.3.2 Wage and Welfare

- **Wage**

First, we compared the difference of mean wage for each type.

The box plot indicates that the larger the company, the higher the salary. And Welch F test indicates the difference of mean wage between companies is statistically significant(p-value<0.05).

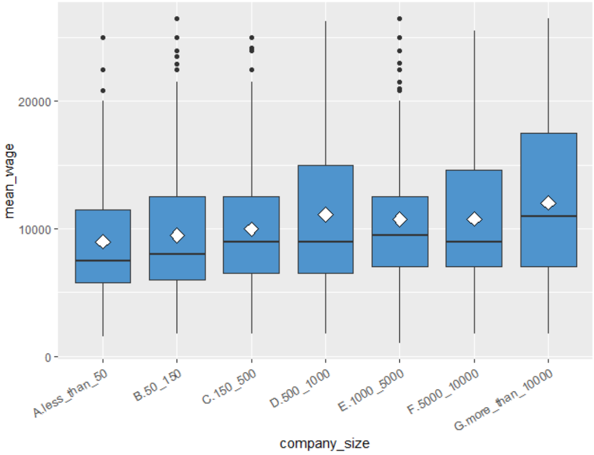

Figure 10: Box plot of company size v.s. mean wage

data	mean_wage and company_size
statistic	39.72198
num df	6
denom df	2423.08
p.value	2.60E-46

Figure 11: Welch's heteroscedastic F test(α=0.05)

The Games-Howell test was conducted to determine which groups have differences and verify that the mean wage is ascending with the company's scale. As can be seen from the table16, some groups with different company sizes but no significant difference in mean wage (such as A&B, C&F). Still, overall, the increase in company size increases the mean wage(estimate is positive) significantly (adjusted p-value <0.05).

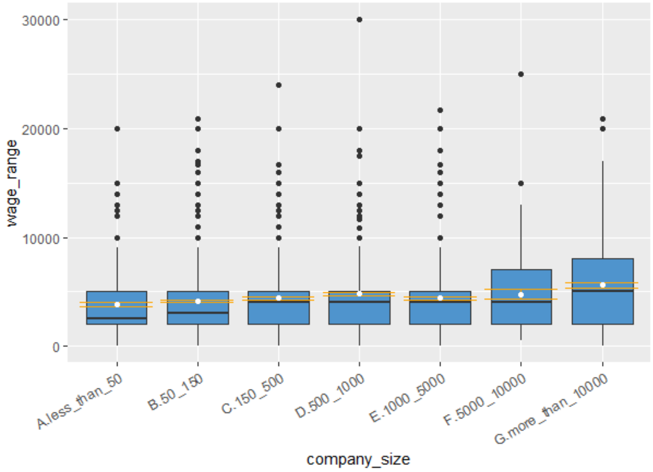

Figure 12: Box plot of company size v.s. wage range

group1	group2	estimate	p.adj	p.adj.signif
A.less_than_50	B.50_150	512.08	1.44E-01	ns
A.less_than_50	C.150_500	1003.87	9.04E-06	****
B.50_150	C.150_500	491.79	1.90E-02	*
B.50_150	D.500_1000	1635.89	0.00E+00	****
C.150_500	D.500_1000	1144.10	1.25E-09	****
C.150_500	E.1000_5000	763.03	1.86E-04	***
C.150_500	F.5000_10000	791.04	1.31E-01	ns
D.500_1000	E.1000_5000	-381.07	4.70E-01	ns
D.500_1000	F.5000_10000	-353.06	9.28E-01	ns
D.500_1000	G.more_than_10000	917.03	5.00E-03	**
E.1000_5000	F.5000_10000	28.01	1.00E+00	ns
E.1000_5000	G.more_than_10000	1298.11	4.24E-06	****
F.5000_10000	G.more_than_10000	1270.09	7.00E-03	**

Table 16: Main results of the game Howell test for mean wage

- **Range of the Wage**

We are also interested in the minimum and maximum wage, so we study the wage range.

From the box plot, we can see that as the company's size increases, the larger the wage spans. Large companies tend to give a more extensive wage range. However, it is worth noting that there are many outliers within small and medium-sized companies, which means that there are many small and medium-sized companies that give a huge wage span for their positions. As can be seen, there are some companies with a wage span of 30,000 Yuan, and many others between 10,000 and 20,000. This means that some small and medium-sized companies set a large wage range, so that the final wage is more flexible.

- **Basic Welfare**

For the welfare column, we use the *strsplit()* function to split the words and count the frequencies. The Top 20 welfare keywords are as follows, and we named these words “Basic Welfare Top 20”.

First, we study welfare based on “basic welfare top20”. For each entry, count the frequency of “basic welfare top20” appearing in the welfare column (variable “wel_count”defined in the Data Processing section). As shown in the figure13(left one), most companies have 1 to 6 basic benefits. From the average and median, the basic welfare of companies of different sizes does not vary much. However, basic welfare has a considerable variation for companies with less than 50 employees and companies with more than 10,000 employees.

rank	item	freq	rank	item	freq
1	五险一金	6349	11	交通补贴	1671
2	绩效奖金	4424	12	通讯补贴	1651
3	年终奖金	3861	13	周末双休	1567
4	专业培训	3722	14	补充医疗保险	1228
5	定期体检	3117	15	免费班车	746
6	员工旅游	3032	16	全勤奖	614
7	餐饮补贴	2993	17	股票期权	556
8	带薪年假	2144	18	出国机会	534
9	节日福利	2082	19	做五休二	487
10	弹性工作	1922	20	补充公积金	372

Table 17: Basic welfare Top 20

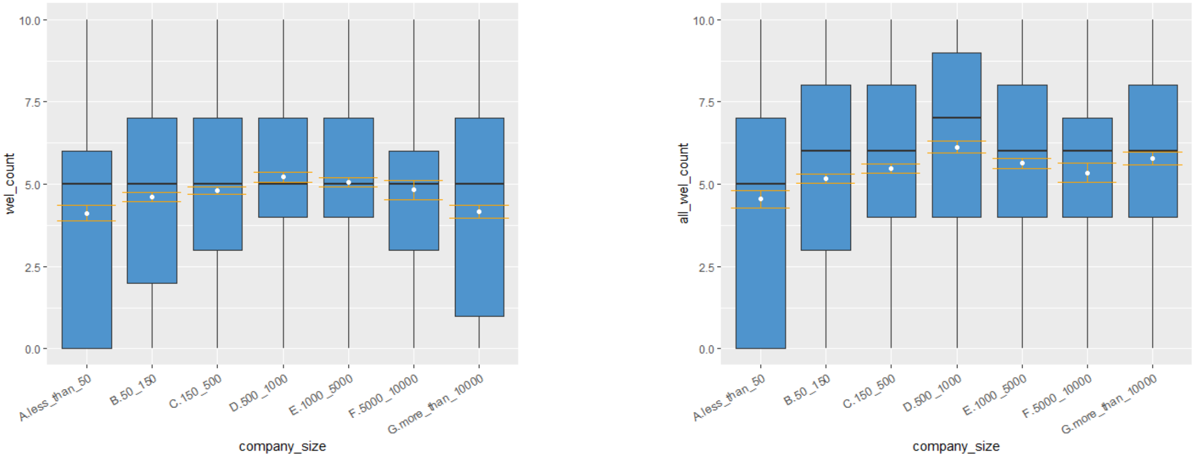

Figure 13: Box plot of welfare frequency (left: basic welfare; right: all welfare)

- **Welfare Counts**

Then, we counted the total number of welfare keywords in each entry. As we can see in the figure13(the right one), it is different from the left one. Companies with less than 50 employees have the lowest welfare and the significant variation. Most of the other companies will give 3 to 8 welfare items. The Game Howell test shows a small but significant difference between large and small companies. In addition, there was no significant difference in benefits between companies E, F, G (with more than 1,000 employees) and companies C (with 150 to 500 employees). Medium-sized companies with 500 to 1,000 employees had significantly higher benefits than other companies.

- **Welfare Contents**

Next, by counting the word frequency and drawing word cloud diagrams, we briefly analyze the content of welfare for companies with less than 50 employees and companies with more than 10,000 employees.

First of all, among the key words of welfare for both sizes of companies, basic welfare such as five insurance and one pension(五险一金), year-end bonus(年终奖金) and performance bonus(绩效奖金) are ranked in the top five. For large companies, the frequencies of welfare such as free meals, free shuttle bus and rental subsidy are higher; for small companies, the frequencies of meal allowance and transportation subsidy are higher. It can be seen that large companies have more subsidies for employees' life.

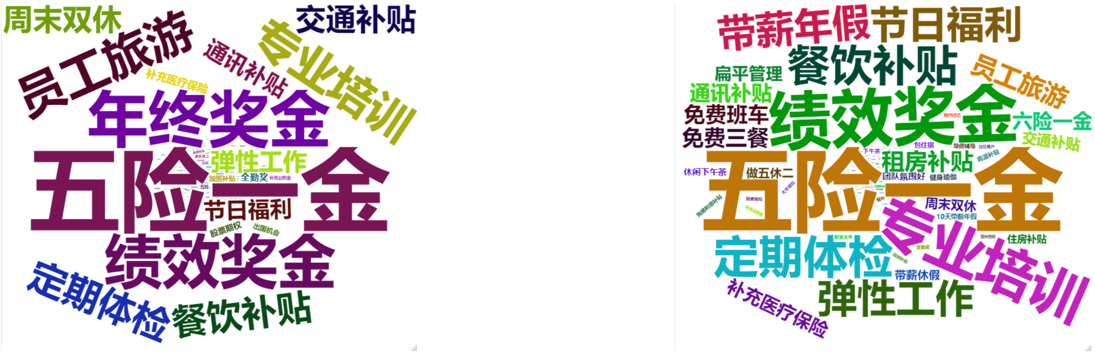

Figure 14: Word cloud of welfare (left-<50; right->10,000)

4.3.3 Demand Distribution

The distribution reflects the demand for data analysts in different-sized companies. Small-to-medium-sized companies with 150-500 employees have the most significant need, reaching about 25%. The demand of companies with less than 50 employees and 5,000-10,000 employees is relatively small, less than 10% of the total sample. In summary, the demand is concentrated in medium-sized companies.

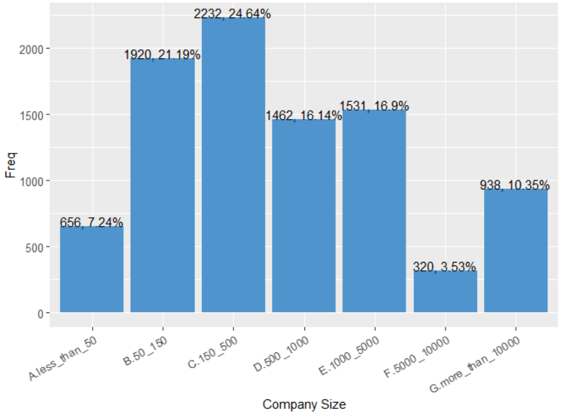

Figure 15: Demand

4.3.4 Requirement for Education Level and Work Experience

- **Education Level**

When studying the differences in the requirements for candidates' education, high school degree, technical secondary degree, and vocational secondary degree were combined; doctor's degree (PhD) and master's degrees were combined into the graduate degree because some sample sizes are too small.

For companies of different sizes, a bachelor's degree is the most popular education requirement. Moreover, as the company's size increases, the proportion of bachelor's degrees increases, and the proportion of the other degrees decrease to a different extent. This indicates that most companies do not require candidates with excessive education, and a bachelor's degree or a technical college degree is enough. However, the acceptance of low education (high school and equivalent degree) is shallow.

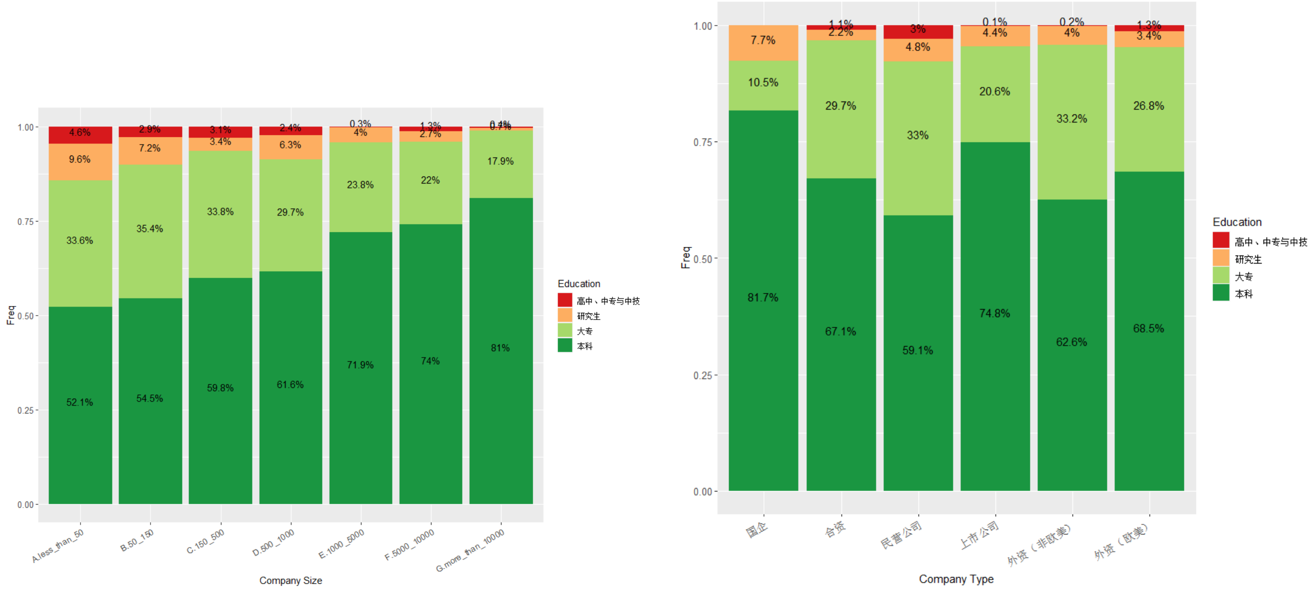

Figure 16: Education requirement (left: company size; right: company type)

It is worth noting that from the perspective of company type, state-owned enterprises, listed companies, and non-European and American foreign-funded companies have basically zero-tolerance towards high school (and equivalent) degrees.

- **Work Experience**

On the whole, the work experience gradually increases with the company's size, but the difference is relatively small. Regardless of large or small companies, most of the required work experience is 0 to 3 years.

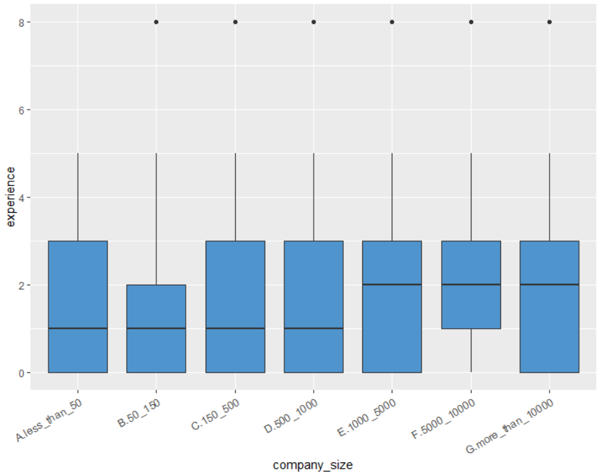

Figure 17: Box plot of experience requirement

Group	Mean	Median	sd
less than 50	1.36	1	1.40
50-150	1.38	1	1.42
150-500	1.60	1	1.58
500-1000	1.63	1	1.63
1000-5000	1.80	2	1.47
5000-10000	2.06	2	1.74
more than 10000	1.86	2	1.60

Figure 18: Statistics for experience

After the Welch test and Game Howell test, as the company's size increases, the company requires more work experience for the candidates. Note that the difference between E (1,000-5,000 people), F (5000-10,000 people), and G (more than 10,000 people) is not statistically significant. This means that these large companies with more than 1,000 employees have similar requirements for work experience.

4.3.5 Summary

During the analysis, large companies with more than 1,000 employees showed a similar nature: high average wage and good welfare. But at the same time, super large companies, especially those with more than 5,000 employees, publish fewer job postings and have stricter education and work experience requirements. These will cause fierce competition when applying for jobs.

For small companies with less than 500 employees, the high demand for talents is significant. Wage levels and welfare are not as good as in large companies. Due to the company's small size, the requirements for work experience and education are also more lenient when recruiting talents.

4.3.6 Additional: Prediction of an Internship

During the analysis, the grouped data often did not have homoskedasticity. We found some large companies with meager salaries for intern positions that did not meet expectations. Thus, we wanted to build predictive models to predict whether a job is an internship position or not by its wage, welfare, size and type of the company, and other variables without knowing the job's title.

Based on the presence of the word "实习生" in the "job_title" or "job_intro" columns, we identified the intern positions in the dataset and then added a column "intern" ("internNO" if it is not an internship).

group1	group2	estimate	p.adj	p.adj.signif
A.less_than_50	B.50_150	0.02	1.00E+00	ns
A.less_than_50	C.150_500	0.25	2.00E-03	**
A.less_than_50	D.500_1000	0.28	2.00E-03	**
B.50_150	C.150_500	0.22	3.53E-05	****
B.50_150	D.500_1000	0.25	6.09E-05	****
B.50_150	E.1000_5000	0.42	0.00E+00	****
C.150_500	D.500_1000	0.03	9.99E-01	ns
C.150_500	E.1000_5000	0.20	1.00E-03	***
C.150_500	F.5000_10000	0.46	1.82E-04	***
D.500_1000	E.1000_5000	0.17	4.20E-02	*
D.500_1000	F.5000_10000	0.43	1.00E-03	***
D.500_1000	G.more_than_10000	0.23	1.50E-02	*
E.1000_5000	F.5000_10000	0.26	1.55E-01	ns
E.1000_5000	G.more_than_10000	0.05	9.80E-01	ns
F.5000_10000	G.more_than_10000	-0.21	4.94E-01	ns

Table 18: Main results of the game Howell test for experience

First, we used logistic regression. Through *step()* and considering the real data, the level, experience, education, company size, company type, and mean wage were used as independent variables for prediction. In the result of logistic regression prediction, the specificity was very low, indicating that the model tends to output a positive result. This is somewhat because the original training set has a significant portion of positive data(internNO). To solve the problem, the training set was down-sampled. The positive data was reduced by 20%, 40%, 60% and 80%, respectively, and the results of the logistics prediction are as follows. It does not reach 0.5 until we down-sampled 40% of the positive data. However, the negative predictive value is declining as a cost.

So, we called the *train()* function in the *nnet* package, and built a feed-forward neural network for prediction. The feedforward neural network was the first and simplest artificial neural network devised. The information moves in only one direction—forward—from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network(Zell Andreas, 1994). Through the results in the table, we can see that the results of each parameter are more satisfactory, and the specificity has improved significantly. Combining the above results, the overall prediction accuracy of the feedforward neural network is better, although the interpretation is poor.

	logistic	down-sample	down-sample	down-sample	down-sample	neural
	regression	(-20%)	(-40%)	(-60%)	(-80%)	network
accuracy	0.9665	0.9669	0.9680	0.9617	0.9407	0.9772
sensitivity	0.9954	0.9938	0.9911	0.9818	0.9528	0.9950
specificity	0.3969	0.4351	0.5115	0.5649	0.7023	0.6260
pos pred value	0.9702	0.9720	0.9756	0.9780	0.9844	0.9813
neg pred value	0.8125	0.7808	0.7444	0.6116	0.4299	0.8632

Table 19: Result of different methods for prediction of internship

4.4 Overall Analysis

4.4.1 Logistic Regression

After backward stepwise, the dependent variable is log(mean_wage) and the independent variables are level, industry_1, experience, education, company_type, company_size, skill_count, wel_count, experience×education, education×skill_count, experience×skill_count, company_size × wel_count. R-square is 0.3965.

- **Skill and welfare both have a positive effect on wage**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
skill_count	0.026757	0.003643	7.346	2.23e-13 ***
wel_count	0.014446	0.005154	2.803	0.00508 **

Table 20: Results of skill_count and wel_count in regression

By mastering a new key skill, the mean wage will increase by 2.67% on average, so it is really important to learn well in the Statistical Computation and Software class. Additionally, more welfare indicates higher wage.

- **Finance, computer science and environment are Top 3 highest wages industries**

The benchmark of industry_1 is Real estate and construction. From the coefficients, we conclude Accounting / Finance / Bank / Insurance industry pays the highest wages and then is Computer / Internet / Communication / Electronics. The profile confirms the reasons why these majors are popular in SUSTech.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
会计/金融/银行/保险	0.186653	0.025296	7.379	1.74e-13 ***
计算机/互联网/通信/电子	0.106699	0.02128	5.014	5.43e-07 ***
能源/环保/化工	0.068212	0.03431	1.988	0.04683 *

Table 21: Partial results of industry_1 in regression

Energy / Environmental protection / Chemical industry is usually considered a low-paid industry, but data analysts can get well-paid wages in this industry.

4.4.2 Random Forest

Before building the random forest, we turn mean_wage from a numerous variable to a categorical one:

	high	medium	low
mean wage	>12500	6500 ~12500	<6500

Table 22: Form conversion

Through trail and error, we conclude the best parameter for random forest is mtry=4 and ntree=1000. The result is as followed, the OOB estimate of error rate is 24.51%, proving that our model recognition is quite accurate:

	high	medium	low
high	1526	493	61
medium	451	3504	547
low	55	613	1809

Table 23: Confusion matrix for mean wage

The Importance of Factors

Using the *importance()* function, we get mean decrease accuracy, which is experience >level >education >company_size >wel_count >skill_count. Therefore, to get a well-paid job, job seekers need to gain wider experience, work in a Tier 1 city, get higher education and work in a big company.

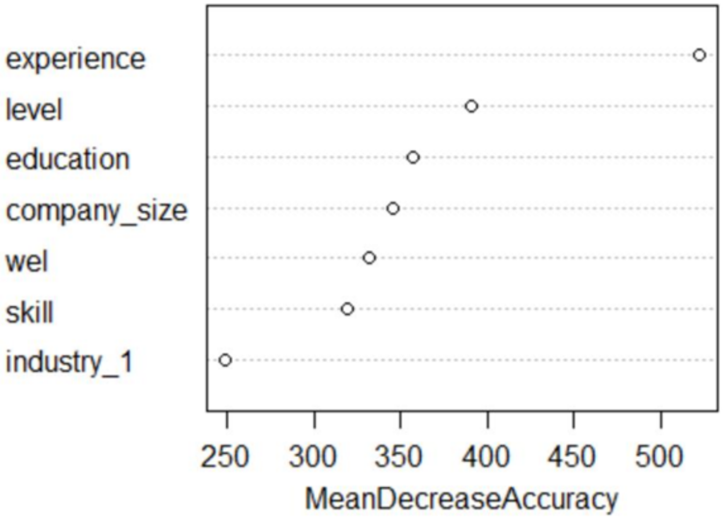

Figure 19: Importance of variables

5 Conclusion

This data analyst recruitment data set provides a unique opportunity to study the wage of data analysts from all aspects.

- Different education has the statistically significant effect on wages, and more skills are required for the bachelor's degree or above. Accumulating experience will greatly help increase wages.
- The Tier 1 cities are better job search areas considering wage and demand. Beijing, Shanghai are optimal. In these cities, computer / Internet/communications/electronics industry demand for data analysis positions are the largest, the skills required includes some large-scale data processing tools.
- For different-sized companies, the employees tend to have high average wage and good welfare. But the candidates may face fierce competition. While for small companies, the high demand gives the job seekers more opportunities and the requirements for work experience and education are more lenient compared to the large companies.

On the whole, the report uses statistical theory to visualize the big data, perform significance tests and give some predictive models, and finally, gives a basic and comprehensive guide for job seekers looking to recruit about data analysis positions.

References

CIACT, China Big Data Industry Development White Paper(2020).

Lee, S., Lee, D. K. (2018). What is the proper way to apply the multiple comparison test?. Korean journal of anesthesiology, 71(5), 353–360. https://doi.org/10.4097/kja.d.18.00242

Joffy Z.,https://rpubs.com/Joffy_Z/DA_analysis.

Zell, A. (1994). Simulation neuronaler netze (Vol. 1, No. 5.3). Bonn: Addison-Wesley.