678final

Qihan Su

2022-12-05

Abstract

According to our career path which is data scientist, the most striking topic is the pay-back of achieving our career goals. What exactly affects the income level of data scientists is a topic that is of great interest to many people.In this project, I focus on the 15 companies with the most statistical data in the data-set as the main object of study. After analyzing the correlation through EDA, I found that the salary of people from different companies with different job titles were affected differently, so I chose to build a multilevel model for stratified analysis.

Introduction

Usually, salary is correlated with the company you work for, the number of hours you work, the number of years you have worked there, and many other factors. In the data science industry the annual salary may vary depending on various job types, for instance, the salary of a product manager in a technology company is definitely different from which is a general data analyst. Also, a data analyst's salary may also be related to the length of time he has been with the company, because the salary is affected by both shares and bonuses, and the share of stock received varies with the time of entry into the company, which can affect the salary level. Lastly, the place where the companies are located may also affect the salary. In order to take all these circumstances in to consideration, I choose to construct a multilevel model with groups of company and job titles.

Method

The data is from the Kaggle website(https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries). The original data has 29 columns with 62642 observations. Due to the large volume of data, I have selected the data of the top 15 companies according to their frequency of appearance in the data set. These companies cover various fields such as technology, finance, and communication, which are of reference significance. On this basis, I removed the columns with weak practical use, and merged the base salary, share valuation and bonus in the original data into a new column representing the total annual salary. And because of the annual salary is high, so I log the salary in order for the further use. I also put the education level, race, and gender in the data as numbers. (In education, high school is 1, college is 2, undergraduate degree is 3, graduate degree is 4, and doctorate is 5; in race, Asian is 1, Latino is 2, two or more are 3, black is 4, and white is 5; in gender, male is 1 and female is 0)

The specific data description are as below.

Data Preparing

column names	explanation
company	Top 15 counts companies
title	job titles
location	City, State, Country
yearsofexperience	Years of working in this industry
salary	Summary of basesalary, bonus and stockgrantvalue
yearsatcompany	Years of working in this company
gender	1 for Male, 0 for Female
cityid	Specific city
Masters_Degree	Highest Degree is Master
Bachelors_Degree	Highest Degree is Bachelor
Docterate_Degree	Highest Degree is Docterate
Highschool	Highest Degree is highschool
Some_college	Highest Degree is some college
Race_Asian	Race is Asian
Race_White	Race is White
Race_Two_Or_More	Having two or more race type
Race_Black	Race is Black
Race_Hispanic	Race is Hispanic
Race	1:Asian,2:Hispanic,3:two or more,4:Black,5:White
Education	1:Highschool,2:College,3:Bachelor,4:Master,5:Phd
state	Different state

Exploratory Data Analysis

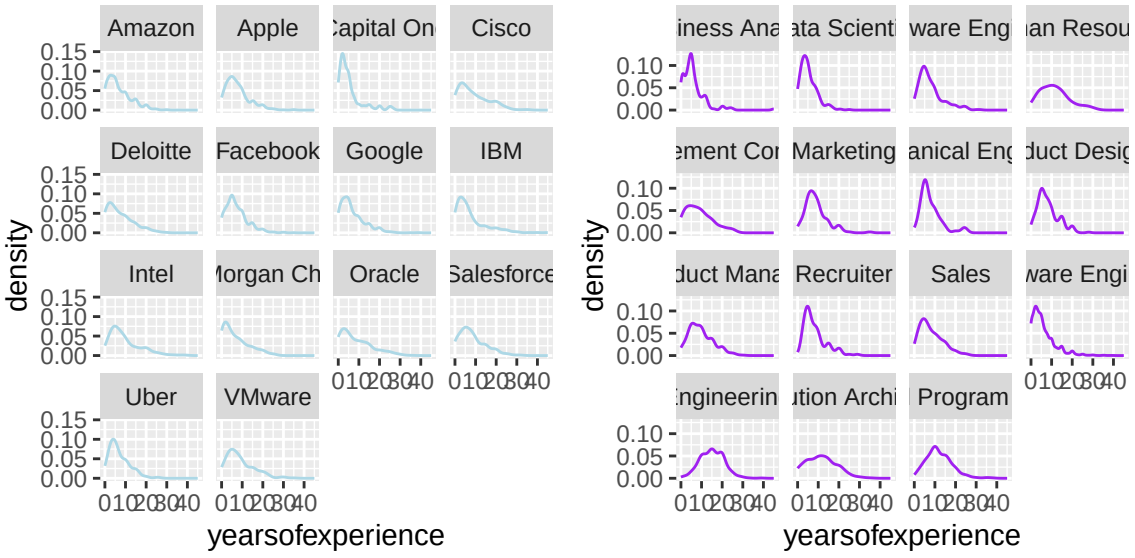


Figure 1: yearsofexperience distribution among different companies and titles.

From the figure 1, we can know that at both the company and job level, it is clear that the more experience you have in this industry, the higher your salary level. However, the correlation between different job positions is not as obvious as at the company level, so we can consider placing the work experience at the job position level in the subsequent modeling.

Figure 2: education distribution among different companies and titles .

The figure 2 above shows the distribution of education between different companies and different jobs. It can be clearly seen that at the company level, education shows a very similar distribution. However, there are more obvious differences in the distribution status of education between different jobs. For data scientists and hardware engineers, the education distribution is more concentrated between 4 and 5, which means that most backend workers tend to be those with master's degrees and PhDs.

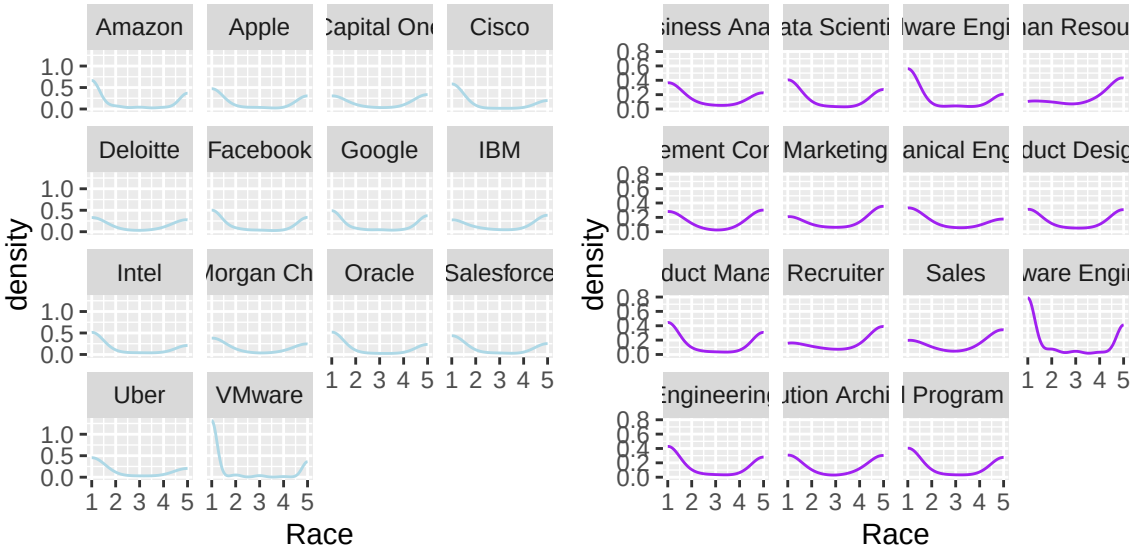


Figure 3: race distribution among different companies and titles.

Since in the data processing stage, 1 represents Asians and 5 represents hundreds. It is more obvious from the figure 3 which is the distribution is relatively similar both at the company level and at the job level.

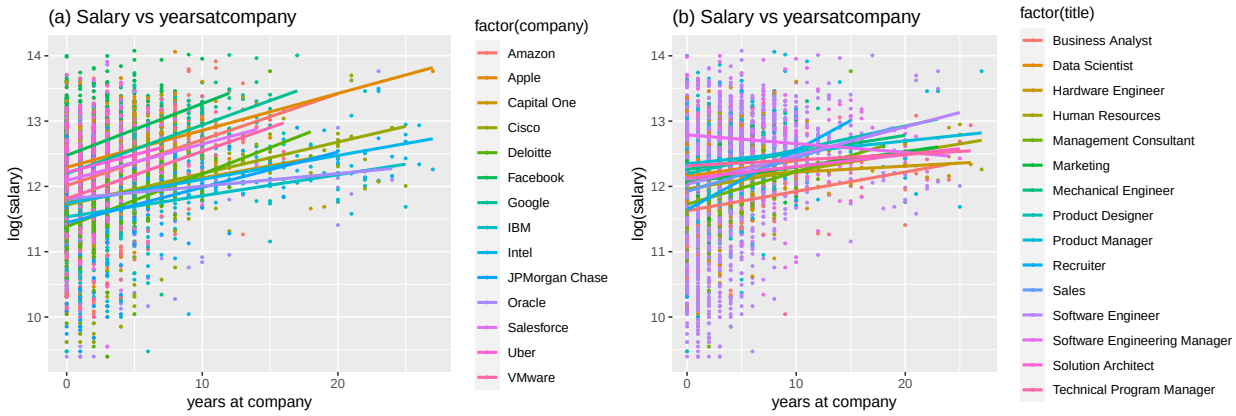


Figure 4: salary vs yearsatcompany.

The figure 4 above shows the correlation between years at the company and salary , both at the company level and at the job title level, with a reasonable increasing. The intercept and slope vary slightly from company to company and from title to title. It seems that although the longer the employee staying at the company, the higher their salary level, it's lightly differ from titles and companies.

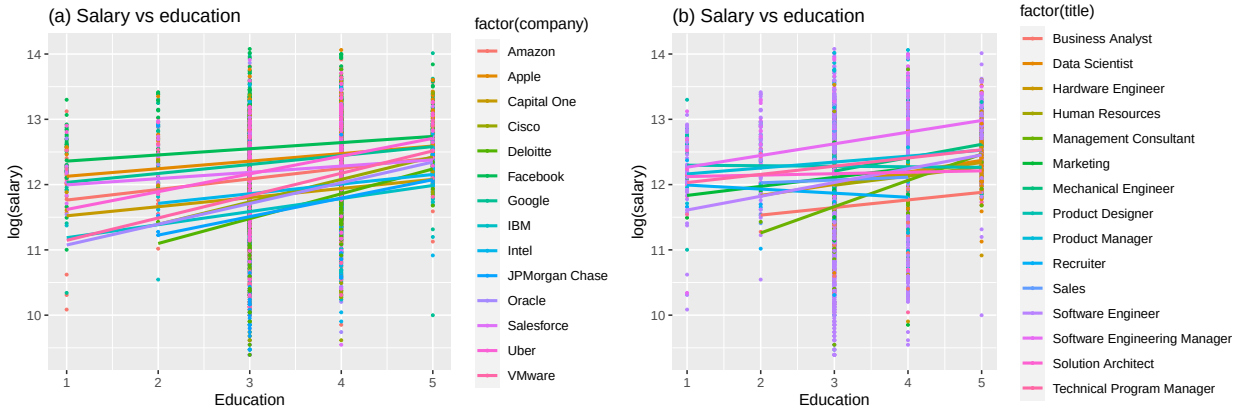


Figure 5: salary vs education.

The figure 5 graphs above show the correlation between the level of education and the level of salary. The two graphs show the correlation between the two at different companies and at different job levels. At either level, there is an overall increasing trend. However, there is a slight difference in the ending and slope at the two different levels. We can also see that for different companies. In technology companies, such as amazon, Google, etc., the degree of impact of education on salary is not as large as in financial industries such as jp morgan, IBM. But at the company level, the overall incremental trend are more obvious. The difference between the intercept and slope is more pronounced at the job title level than at the company level.

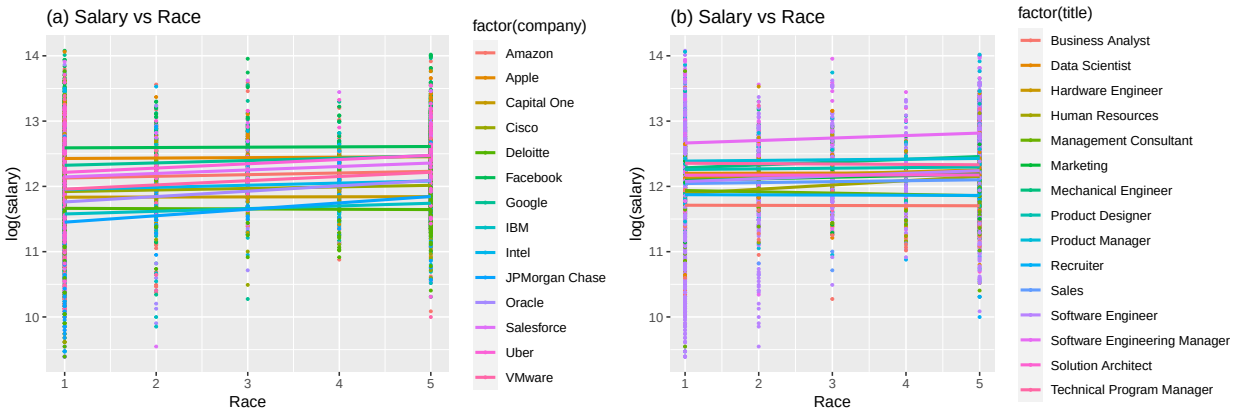


Figure 6: salary vs race.

The correlation between race and salary level is not particularly clear, but the graph shows that among the company level, IBM and JP Morgan show an upward trend, while other companies tend to flatten out. in the race data, from 1-5 represent Asian, Brazilian, multi-racial, hei re. at the job level, there is a slight difference in the intercept and slope

Correlation of Data

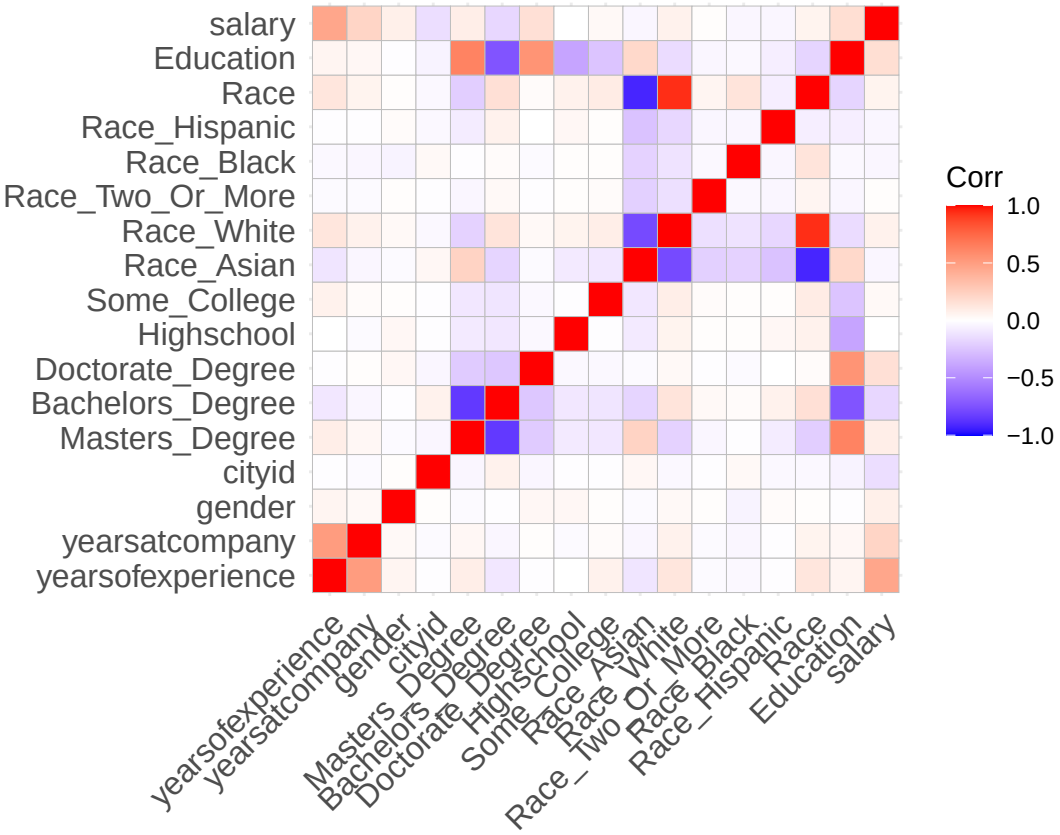


Figure 7: correlation plot

From the figure 7 we can see that there is a strong positive correlation between salary level and years of experience in the company as well as work experience. Secondly, there is a slight positive correlation between gender and salary, and since men are 1 and women are 0, it can be found that the overall salary level is higher for men than for women. In the education category, education shows a relatively significant positive correlation with salary level, and since Ph.D. is a value of 5, it can be seen that having a Ph.D. has a relatively large impact on salary level, but having a bachelor's degree or not does not have a large impact on salary level. Within the race category, it is more evident that the strongest correlation between whites' and salary level can be seen. From the correlation analysis, only the years at the company was retained in the subsequent modeling because the yearsofexperience have strong correlation with yearsatcompany, so we drop the yearsofexperience.

The correlation chart help me to decide the variable which could be used in the later multilevel model constructing.

Model Fit

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.142e+01	1.179e-01	2.573e+01	96.925	< 2e-16 ***
yearsatcompany	4.543e-02	5.399e-03	1.335e+01	8.415	1.06e-06 ***
Education	1.384e-01	1.723e-02	7.719e+00	8.036	5.17e-05 ***
Race	3.808e-02	5.411e-03	9.104e+00	7.037	5.73e-05 ***
cityid	-1.215e-05	7.217e-07	8.137e+03	-16.839	< 2e-16 ***

The chart above is the summary of the fixed effct.

	(Intercept)	yearsatcompany	Race
Amazon	0.12	0.01	0.00
Apple	0.39	0.01	-0.02
Capital One	-0.03	-0.02	-0.01
Cisco	-0.19	-0.01	0.00
Deloitte	-0.46	0.01	0.00
Facebook	0.61	0.03	-0.02
Google	0.29	0.02	-0.01
IBM	-0.43	-0.02	0.01
Intel	-0.18	-0.01	0.00
JPMorgan Chase	-0.44	-0.01	0.22
Oracle	-0.14	-0.03	0.02
Salesforce	0.20	0.01	-0.01
Uber	0.35	0.00	0.00
VMwre	-0.08	0.01	0.01

	(Intercept)	Education
Business Analyst	-0.19	-0.03
Data Scientist	-0.09	0.02
Hardware Engineer	-0.07	0.01
Human Resources	-0.16	-0.02
Management Consultant	-0.06	0.07
Marketing	-0.17	0.02
Mechanical Engineer	-0.06	0.00
Product Designer	0.22	-0.04
Product Manager	0.23	0.00
Recruiter	-0.24	-0.05
Sales	-0.01	-0.02
Software Engineer	-0.12	0.04
Software Engineering Manager	0.40	0.03
Solution Architect	0.26	-0.04
Technical Program Manager	0.06	0.01

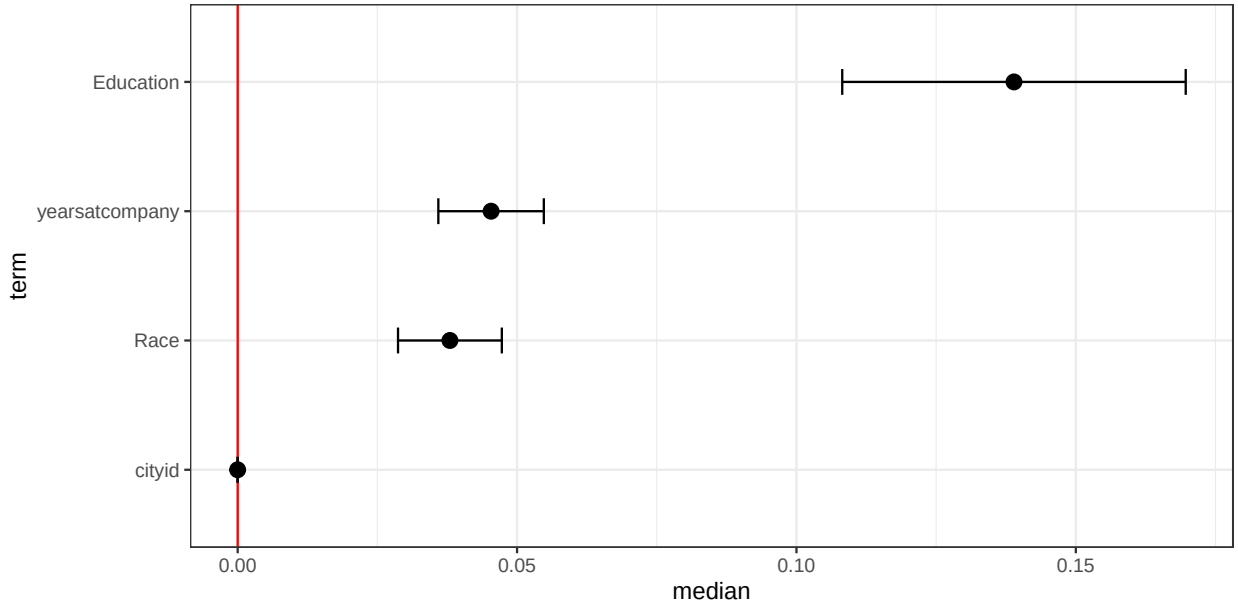
Result

Interpretation The fixed model is below:

$$log(salary) = 11.42 + 0.45 \times yearsatcompany + 0.14 \times Education + 0.38 \times Race$$

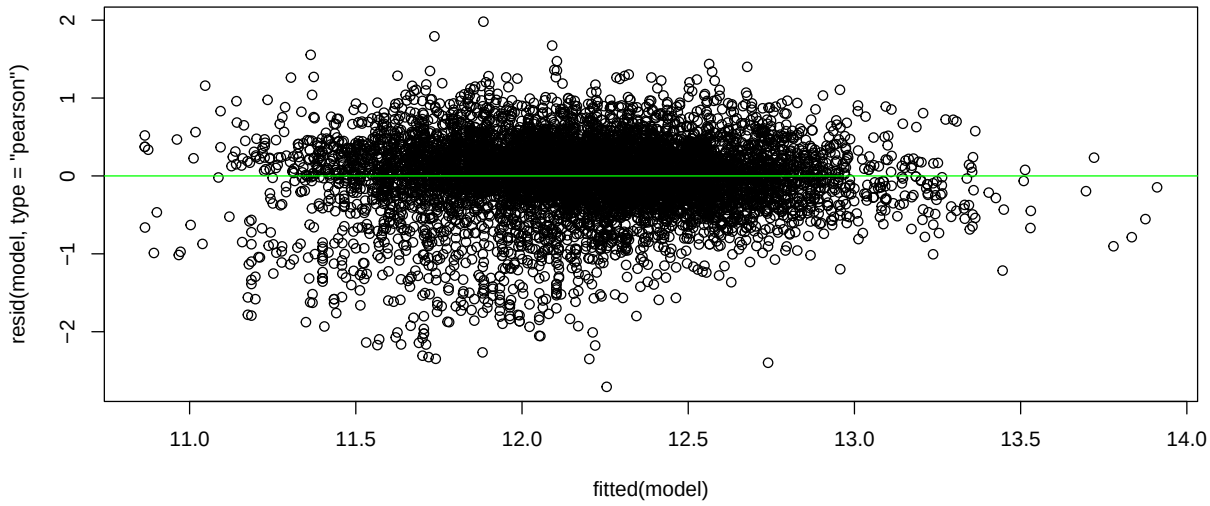
Take the random effect into consideration, and choose the business analyst from the company Amazon as an example.

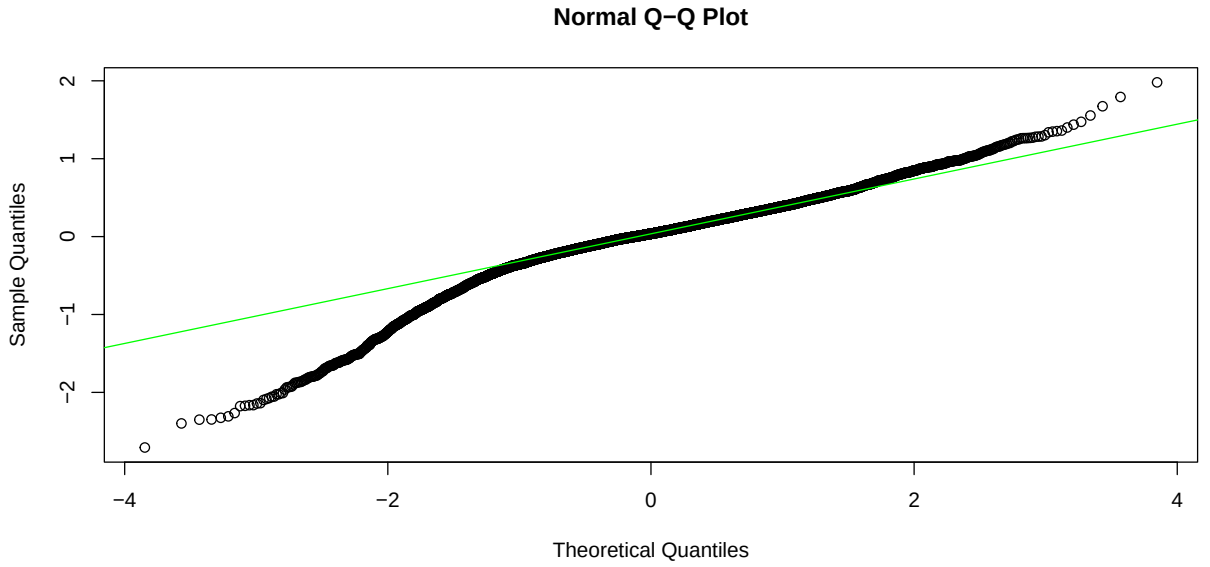
$$log(salary) = 11.4 + 0.46 \times yearsatcompany + 0.11 \times Education + 0.38 \times Race$$



According to the fixed model, it can be seen that the salary was 11.42 when yearsatcompany, Education and Race at their average level. Each year the employee stay at this company, their salary will expectedly grow 0.45, and every level increase in the empoyee's education level will contribute 0.11 increase to their salary level. For the Race, it shows that there's 0.38 positive impact on salary level, I reckon that this was beacuse Whites population is the largest portion in the data-set. When taken the random effect into consideration, it shows that there is a 0.02 difference between different companies and job titles. As we can see, there is a 0.01 difference between salary and different years at companies by different companies, and there is a 0.03 difference can be seen between the Education level and salary by different job titles. However, for Amazon, there is no difference between Race and salary on the company level.

Model Checking





In order to check the model fit, I draw the Q-Q plot to check, from the plot showing above, the model seems to fit well.

Conclusion

In this article, I select a data-set of salary levels in data science and stem majors to explore the factors that affect salary levels. The analysis is also conducted at both company and job levels using a multilevel model based on data characteristics. The model concludes that the length of time working in the company, race type, and education level all show positive correlation with salary level. In terms of job experience, there is a significant increasing trend of higher salary level with longer working experience in the specific company. In terms of race, in this paper, value 1 is used for Asian, value 2 for Hispanic, value 3 for both and more races, value 4 for Black, and value 5 for White. I found it interesting to note that Asians are more likely to work in back office jobs such as software engineers, mechanical engineers, and hardware developers, but less likely to work in consulting, marketing, and human resources positions. In terms of education, the highest number of people in all positions are bachelor's and master's degree. And it can be found that for companies in the financial field such as IBM and JP Morgan Chase, the distribution of education will be higher than that of technology companies such as Amazon and Google. Secondly, the variable of city code is not subsumed into the company as well as the title consideration, but I have compiled the top 15 cities by processing the data. The most people engaged in the field of data science are distributed in Seattle, San Francisco, New York, Menlo Park, Cupertino, Mountain View, Sunnyvale, Austin, Bangalore, San Jose, Palo Alto, Washington, Santa Clara. Nine of the top fifteen cities in the country in terms of the number of practitioners are located in California. This should be due to the fact that many large technology companies are located in California. For example, Seattle is ranked first because Amazon's headquarters is located in Seattle. And memlo park ranked 4th, also because it is the headquarters of Facebook.

Discussion

In this article, I consider the factors that influence salary levels in the data science industry on two levels and analyze these factors in groups. However, there are still some limitations, for example, in the aspect of location, I did not explore in great detail the extent to which location has a shadow line on the level of salary. Secondly, I think the model can make model predictions, such as what combination will be more conducive to career development and achieve a more satisfactory salary, thus helping to provide a better prediction for people who are about to enter the industry.

Appendix

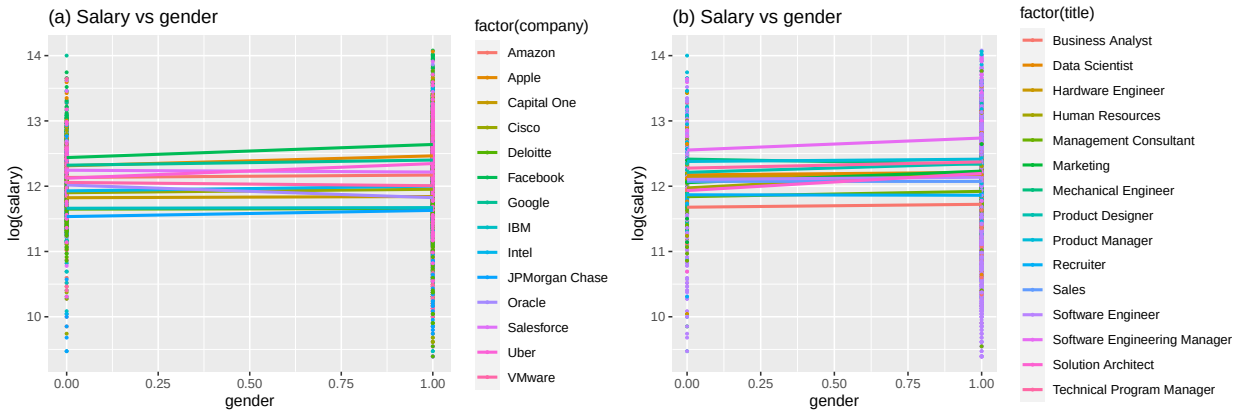


Figure 8: salary vs gender.

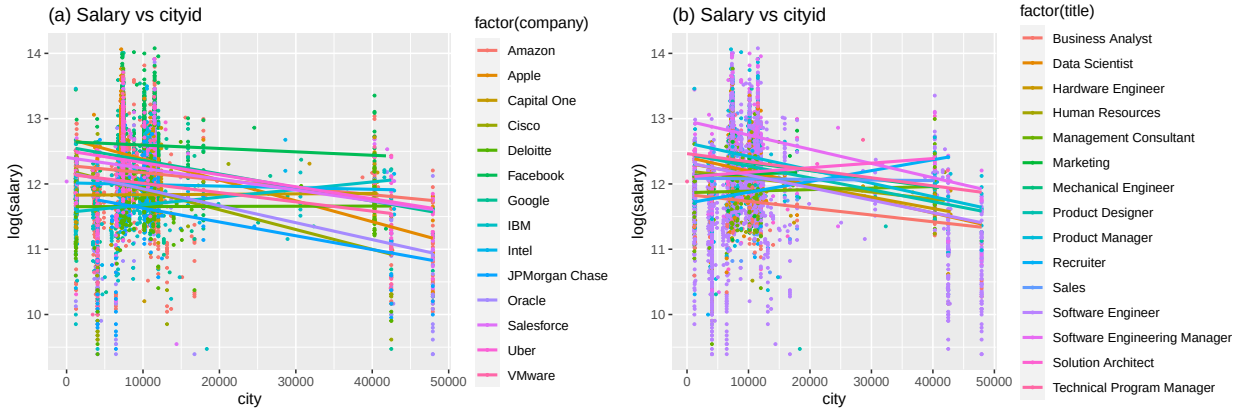


Figure 9: salary vs city.

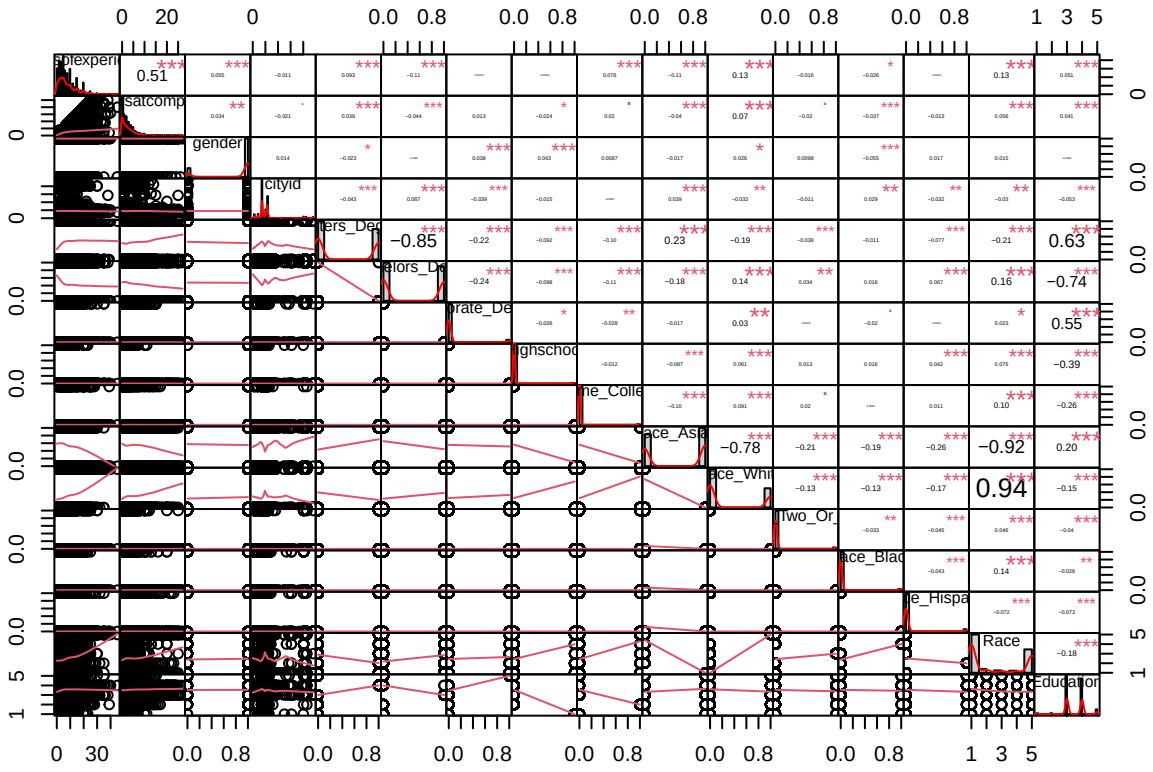


Figure 10: Correlation Matrix