MA678 Midterm Project

Shicong Wang

11/13/2021

Abstract

Employment has always been a source of pressure for graduates. Everyone is eager to find the most suitable job for them, holding a satisfactory salary in a favorite company and handy position. However, graduates are often confused by complicated information. They need to compare salaries in multiple companies and positions, and they are also screened by the various requirements of different companies. As a result, making analysis in the factors effecting salaries attaches vital importance. Graduates are supposed to make sense so that they can try to meet the requirements before graduating, and given to their conditions choose the most suitable job after graduation.

Introduction

The factors that determine salary are complex. It may be related to the employee's personal characteristics, such as the employee's education level, work experience, and even gender and race. Additionally, it is with respect to the company itself like the location of the company and the company types. Even in the same company, the salary of an employee is closely depended on his position and rank. Therefore, the analysis of factors affecting salary should be comprehensively considered from multiple aspects and perspectives. Although it seems natural that a PhD may worth higher salary than a master degree, or a experienced staff is more welcomed than a graduation. Nevertheless, when it comes to the comparison between the salaries of an Asian male graduation with a PhD in software engineering at Apple Inc and a white girl working five years with a master's degree in data science at Google, things can be confused.

As a result, a simple linear regression model cannot solve this problem containing various types of dependent variables. In the following steps, I will build a multilevel model to research the data set whose data for participants are organized at more than one level.

Method

Data Cleaning and Processing

The data set I choose can be downloaded on Kaggle: Data Science and STEM salaries. The data set contains more that 62,000 STEM salary records from 1085 top companies all around the world, and involves useful information such as education level, compensation (base salary, bonus, stock grants), race, and more, which serve as the dependent variables in the model.

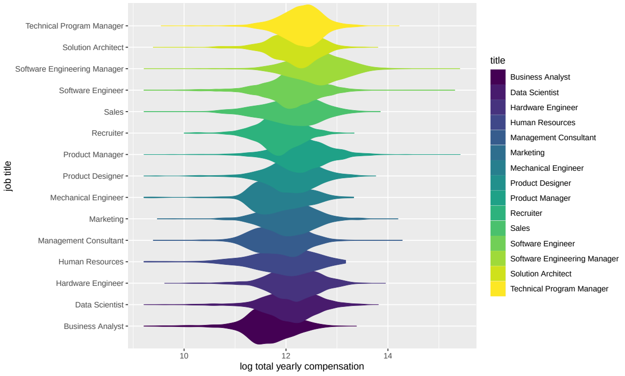
Since the data set came from a survey, the surveyors did not pay attention to the uniformity of the company name format when filling in the questionnaires, for instance, Jp Morgan showed in different format like "JpMorgan", "Jpmorgan", so firstly I tried to make sure that one company name is presented in one form. Secondly, I separated the columns with abundant information to guarantee that the information in the each column is in more details. Thirdly, I dropped the default values in the data set. Finally, I selected some columns and mutated them into new sub-data-sets to make comparison in different dimensions. Here are some explanations of columns:

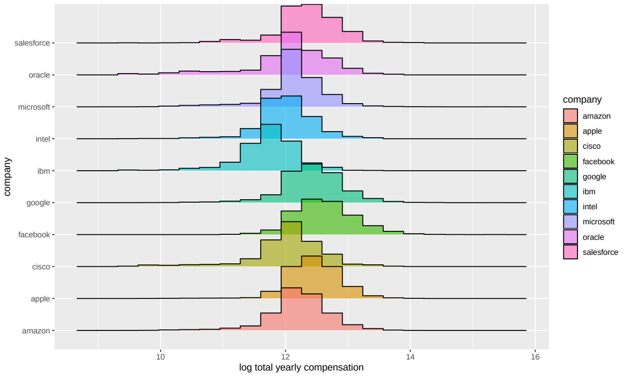
column names	explanations
title	The Specific position in companies
totalyearlycompensation	Cumulative value of one year's salary
level	The ranks within the companies
yearsofexperience	How long is the staff works
yearsatcompany	How long is the staff in this company
states	The states where the companies are located
edu_level	Five levels according to acedemic degree
work_experience	Four levels according to the years of experience

Exploratory Data Analysis

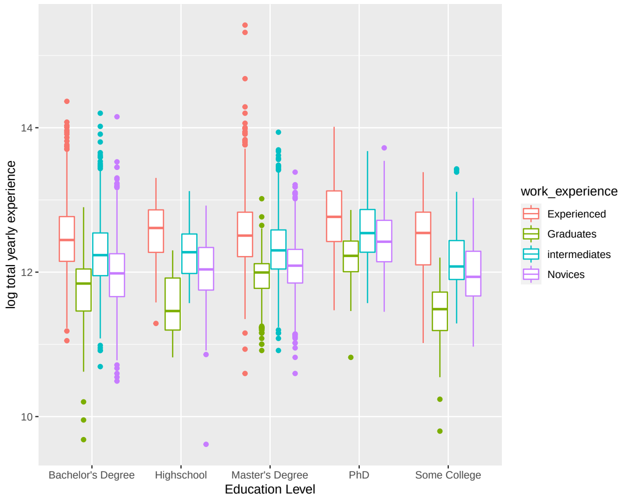
As the factors taking into consideration are multiple, it is unrealistic to present these factors in one picture, so I make efforts to draw a series of plots to show the effects of different factors on salary. Due to the large difference in magnitude among the variables, most of the points will accumulate at the bottom of the image. As a result, I use "log(totalyearlycompensation)" to substitute "totalyearlycompensation". Maybe the image is not as intuitive as before, but the distribution of points can be seen more clearly.

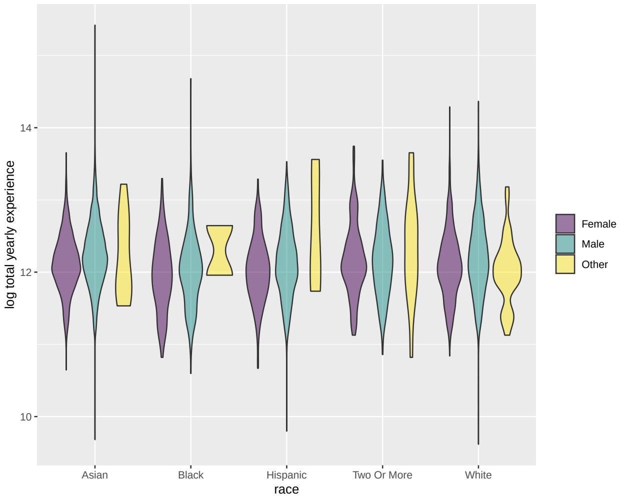
Here some plots to see if there is correlation among job titles and companies with total yearly compensations.



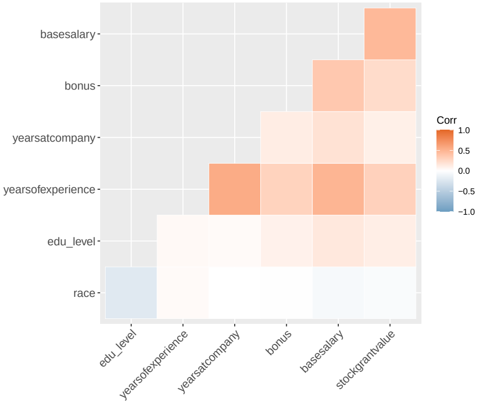


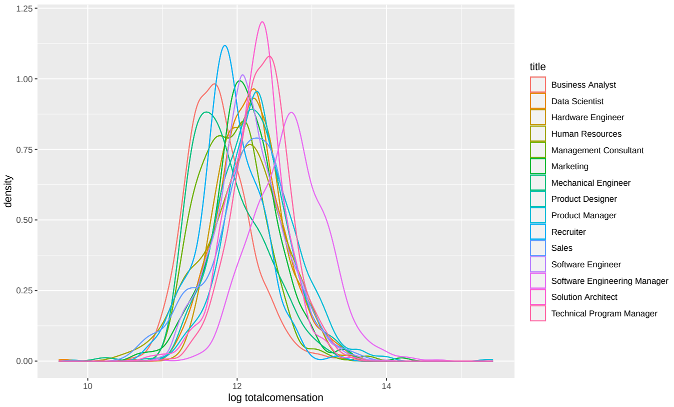
The plots above respectively show the relationship between job titles, companies and total yearly compensations. Apparently, both of them exert effect on total yearly compensations. When it comes to job titles, software engineering manager and product manager seems enjoy better salary treatments. Besides, when it comes to companies, Google, Facebook and Salesforce obtain higher employees' salaries, whereas IBM's salaries are less attractive. When analyze staff salaries are related to which factors, both can be used as the basis for grouping staff.





The third plot shows the correlations among total yearly compensation, years of experience and education level. The result was unexpected. Education level and years of work experience affect the salary level to a certain extent, but not exactly as we imagined. Generally speaking, a Ph.D. has a significant advantage in salary. For other education level, perhaps a higher degree has certain advantages when first enter the job, but with the accumulation of work experience, such advantages become less obvious. Of course, more work experience does mean more salaries in this plot. The forth plot make comparisons of total yearly compensation among race and gender.To relief, neither of these factors have a significant impact on salary.It can be seen that male employees have certain advantages in the high-income range, but generally speaking, the median and mode of male and female employees' income are not much different.





Finally, make preliminary preparations for the establishment of the model. Use correlation plot to observe the correlation between the selected independent variables. At the same time, according to the distribution chart, it can be seen that the salary distributions of different job titles are different, which will be selected as the basis for grouping. Since the number of company in the data set is 1085, which is too large for grouping, so it's more proper to choose title.

Model Fitting

And to see the fixed effects below, all variables are significant at alpha = 0.05 level.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	11.563570	0.077259	16.75	149.673	6.34e-13 ***
yearsofexperience	0.042733	0.002434	24.53	17.557	2.10e-15 ***
yearsatcompany	-0.017426	0.001073	12.21	-16.243	2.99e-06 ***
gender	0.065774	0.013228	10.98	4.972	1.21e-09 ***
race	-0.018349	0.003658	10.98	-5.016	1.21e-09 ***
edu_level	0.103877	0.013031	10.98	7.972	1.21e-09 ***

Result

Model Coefficients

Given the model fit above, we can conclude this formula:

$$y = 11.56 + 0.043x_1 - 0.017x_2 + 0.066x_3 - 0.018x_4 + 0.104x_5$$

Let $x1 = yearsofexperoenve$, $x2 = yearsatcompany$, $x3 = gender$, $x4 = gender$, $x5 = educationlevel$, $y = log(totalyearlycompensation)$.

Since the magnitude value of total yearly compensation, in order to avoid the coefficient of the regression formula being too large, use log instead. The education level and ethnicity are assigned separately and added as independent variables to the regression model.So the formula contains two continuous variables and three discrete variables.Perhaps the coefficient of years at company is a bit strange, cause according to common sense, as the working time in the company increases, the rank will increase to a certain extent, which means salaries are supposed to rise, and the coefficient should be positive.However, the few years of working in the company does not mean that the staff is a novice. Some experienced employees will choose to quit, which means that even if they don't stay in the new company for a long time, they still get a good salary with their rich experience.

As for different job title, the degree of influence on each independent variable is different.I choose the representative job titles below:

Job Title	(Intercept)	yearsofexperience	gender	race	edu_level
Software Engineer	11.78	0.041	0.042	-0.023	0.111
Data Scientist	11.50	0.043	0.056	-0.020	0.159
Business Analyst	10.95	0.048	0.057	-0.012	0.080
Hardware Engineer	11.29	0.045	0.049	-0.021	0.133
Software Engineering Manager	13.53	0.024	0.092	-0.022	0.080
Recruiter	11.91	0.040	0.057	-0.008	0.068

Maybe the distinction of the coefficients can be explained in this way: Let's take education level as example. It seems that education level have the most significant impact on data scientist and the least impact on recruiter. Simply analyze the data set and findings can be interesting: that is, in the recruiter, most of the staff are undergraduates, while in data scientist, the staff's education level are mainly focused on undergraduates, masters, and PhD degrees, so they can compare the impact of education level on salary. Moreover, when it comes to years of experience, the coefficient of software engineering manager is lowest since majority of software engineering manager all have extensive work experience and have worked for more than 8 years, as a result, the decisiveness of work experience is greatly reduced.

Model validation

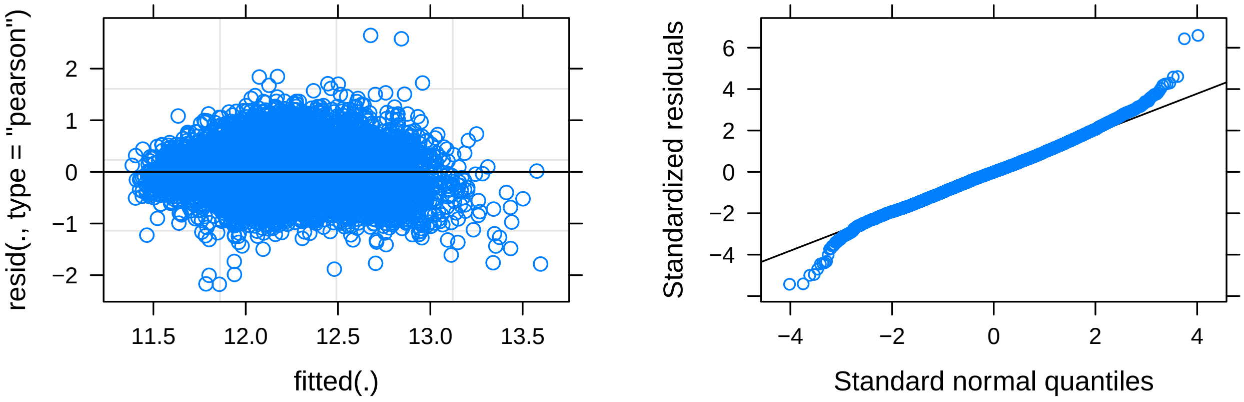
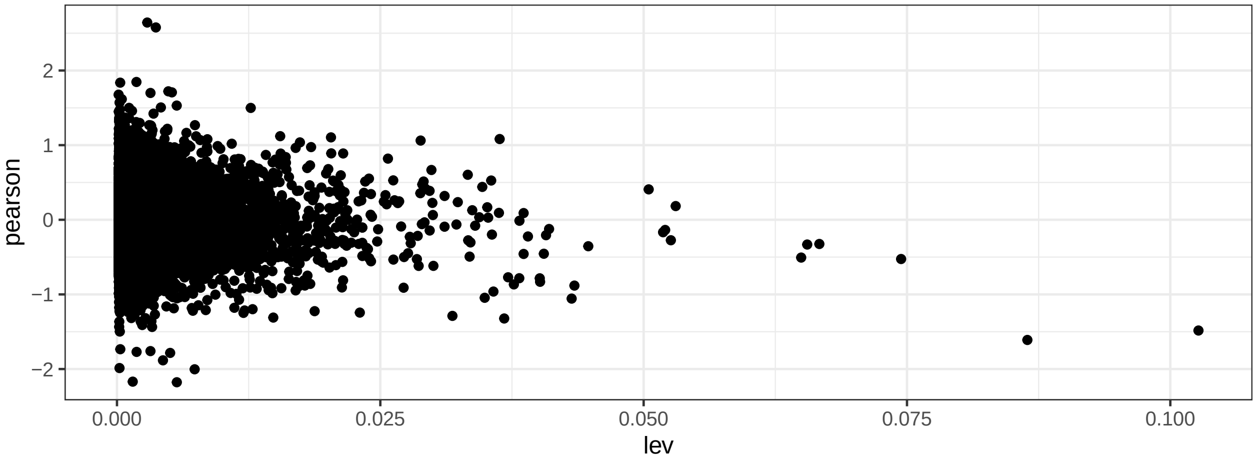


Figure 1: Residual plot and Q-Q plot.



From the Residual plots in Figure 4 we can see that the mean of residuals is almost 0, which means this plot makes sense. As for the Q-Q plot ,the majority of dots are on the lines so the normality is good except for some of samples may distribute as a fat tail. Figure 5 shows that there are not obvious leverage point.

Discussion

In a sense, this model is reasonable.Relatively speaking, a person with many years of work experience and a higher degree is indeed more likely to get a higher salary. At the same time, in the sample of this data set, men are slightly better than women in both the number and salary in the STEM field. But this does not mean that women will be inferior to men in the field of STEM.

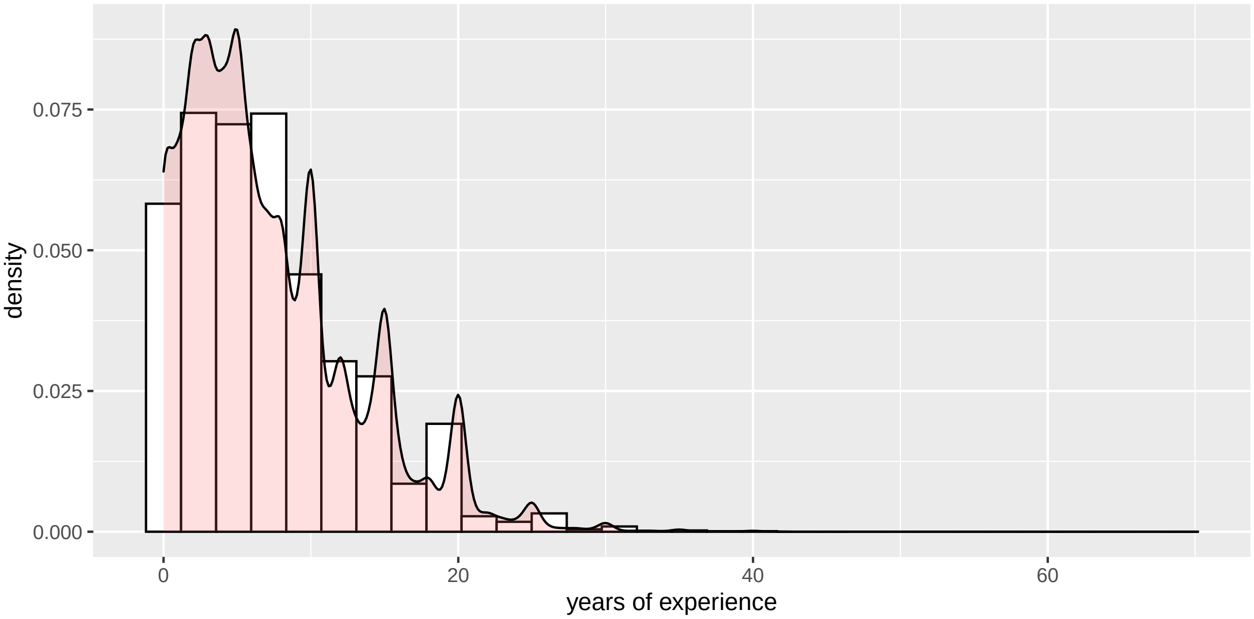
In fact, there are many factors that can be taken into consideration in this analysis. For example, the salary level of different companies will be different, but because there are 1085 companies in the table, it is difficult to distinguish them according to their types, because some companies are comprehensive companies involving multiple fields. At the same time, analyzing correlation of the salary and state, it can be found that the salary level of each state will also be different. However, there are too many factors involved in the salary difference between state and state, such as policy factors, economic development level, etc., and it is difficult to conduct further analysis only through the factors in the table.

Appendix

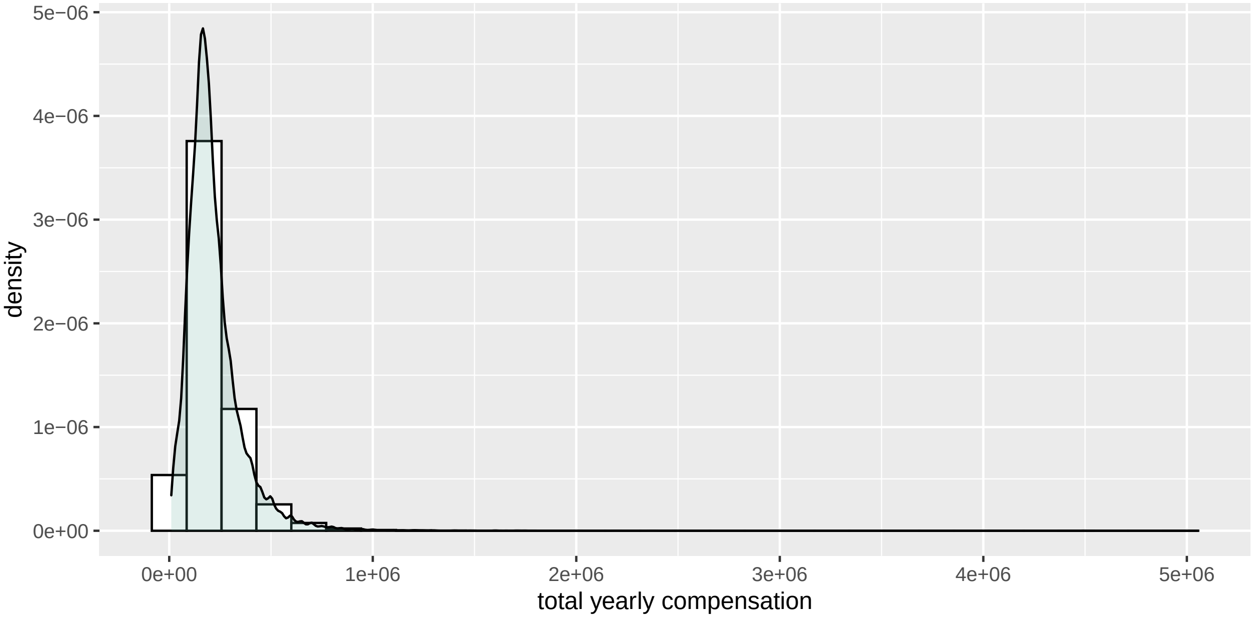
More EDA

Density plot

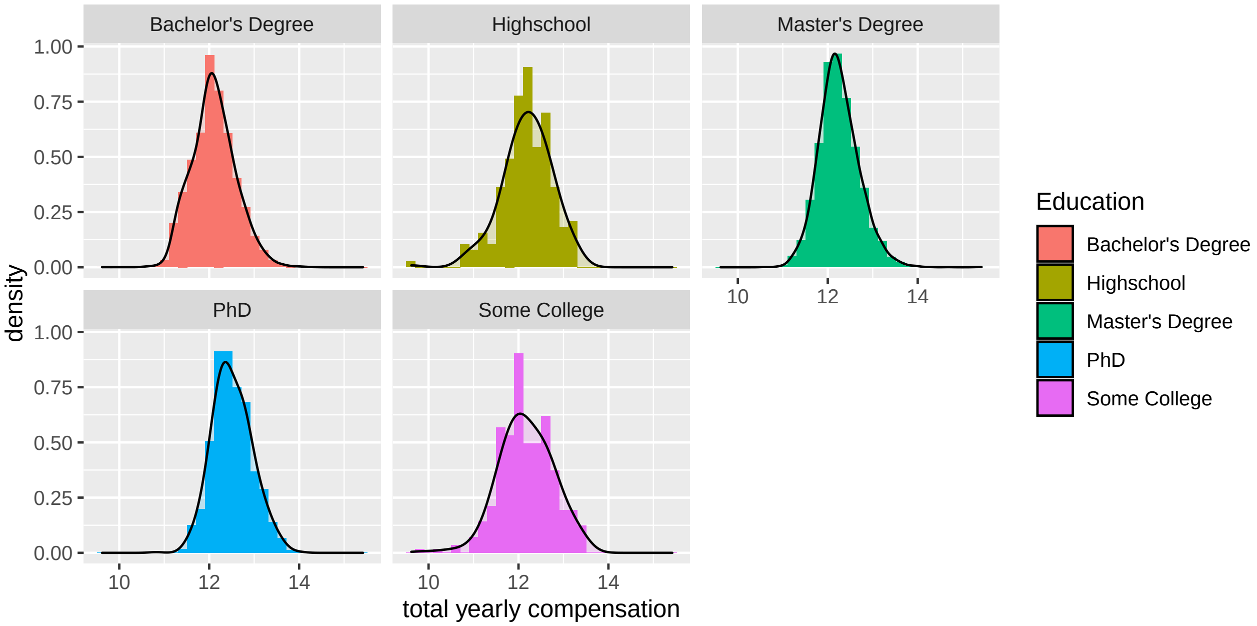
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

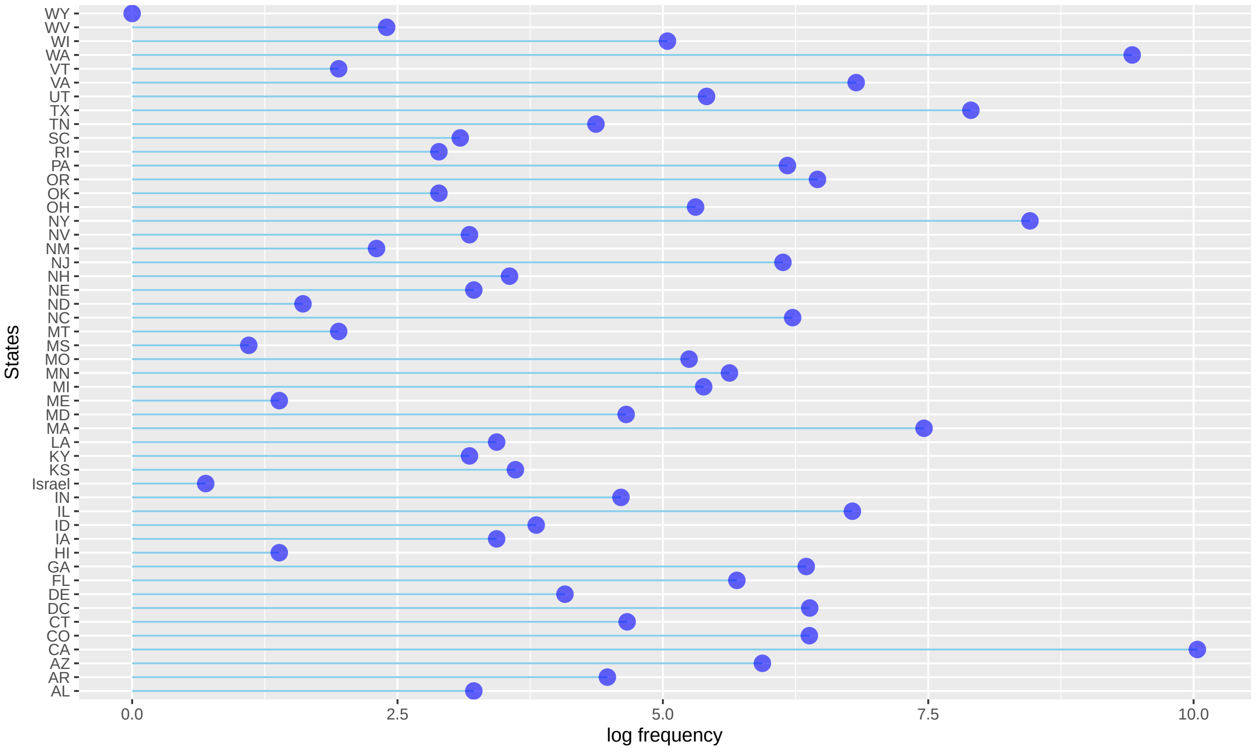


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



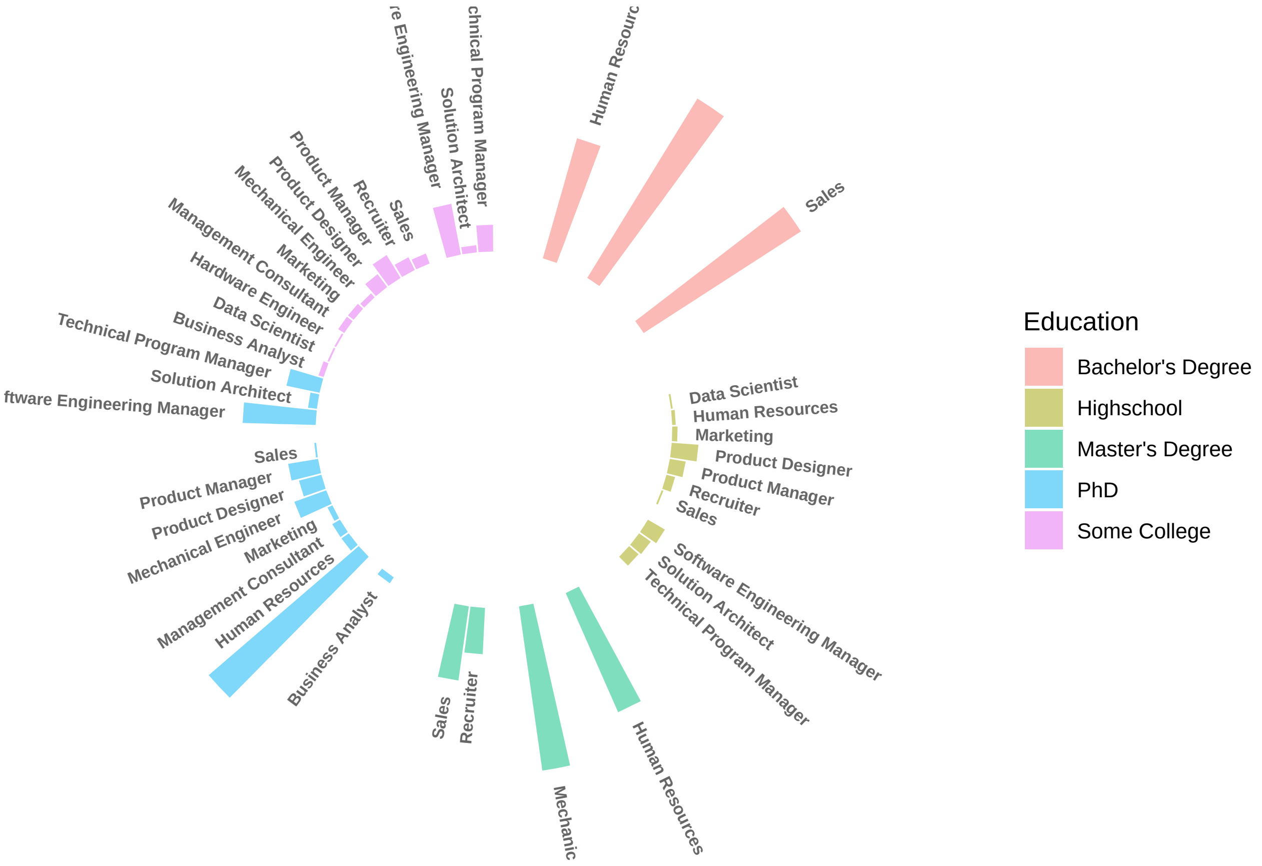
Lollipop

Count of company in each state



Circular Bar Plot

Distribution of job title in each educaiton level



Full Results

Random effects of model

```
## Warning: NAs introduced by coercion
```

```
## boundary (singular) fit: see ?isSingular
```

```
## $title
##                              (Intercept) yearsofexperience (Intercept)
## Business Analyst             -0.07560963      0.0037164013           0
## Data Scientist               -0.02287333      0.0011242809           0
## Hardware Engineer            -0.06910213      0.0033965415           0
## Human Resources              -0.02279931      0.0011206428           0
## Management Consultant        -0.06651095      0.0032691788           0
## Marketing                    -0.13937120      0.0068504410           0
## Mechanical Engineer          -0.18835533      0.0092581329           0
## Product Designer             -0.07658912      0.0037645457           0
## Product Manager              -0.07067375      0.0034737906           0
## Recruiter                     0.11331217     -0.0055695751           0
## Sales                        -0.24748757      0.0121646295           0
## Software Engineer             0.01738696     -0.0008546121           0
## Software Engineering Manager  0.33311389     -0.0163733765           0
## Solution Architect            0.25133398     -0.0123536905           0
## Technical Program Manager     0.26422530     -0.0129873308           0
##                                     gender   (Intercept)          race
## Business Analyst             -0.0119039440 -1.968571e-05  2.909096e-06
## Data Scientist               -0.0096090016  5.602780e-06 -8.279622e-07
## Hardware Engineer            -0.0201533873 -2.630487e-06  3.887256e-07
## Human Resources               0.0095602004 -4.003487e-06  5.916235e-07
## Management Consultant         0.0157889897  3.172414e-05 -4.688099e-06
## Marketing                     0.0106978902 -8.502280e-06  1.256442e-06
## Mechanical Engineer          -0.0031577916 -7.290636e-06  1.077389e-06
## Product Designer              0.0254322610  2.025083e-06 -2.992607e-07
## Product Manager               0.0056941427  4.956895e-05 -7.325152e-06
## Recruiter                    -0.0047867897 -2.862111e-05  4.229542e-06
## Sales                        -0.0009068594  1.526812e-05 -2.256278e-06
## Software Engineer            -0.0248688868  2.332384e-05 -3.446727e-06
## Software Engineering Manager  0.0236353901 -1.842394e-05  2.722635e-06
## Solution Architect            0.0118784912 -3.364967e-05  4.972648e-06
## Technical Program Manager    -0.0273007048 -4.705591e-06  6.953783e-07
##                              (Intercept)    edu_level (Intercept)
## Business Analyst              0.05129376 -0.022862344 -0.23272279
## Data Scientist               -0.12392527  0.055235221  0.08395396
## Hardware Engineer            -0.05944436  0.026495179  0.06641260
## Human Resources              -0.02411662  0.010749115 -0.16940100
## Management Consultant        -0.09093022  0.040528864 -0.19411430
## Marketing                    -0.03646351  0.016252295  0.02782273
## Mechanical Engineer          -0.06689318  0.029815223 -0.05103377
## Product Designer              0.03762267 -0.016768949  0.15978026
## Product Manager               0.02867766 -0.012782031  0.22222709
## Recruiter                     0.07782263 -0.034686632 -0.26566695
## Sales                        -0.02979546  0.013280250  0.20287640
## Software Engineer            -0.01414694  0.006305488  0.12787810
## Software Engineering Manager  0.06303529 -0.028095708  0.24540122
## Solution Architect            0.09141620 -0.040745475 -0.17776404
## Technical Program Manager     0.09584735 -0.042720498 -0.04564951
##                              yearsatcompany
## Business Analyst               0.0077273704
## Data Scientist                -0.0027876229
## Hardware Engineer             -0.0022051762
## Human Resources                0.0056248220
## Management Consultant          0.0064454071
## Marketing                     -0.0009238312
## Mechanical Engineer            0.0016945349
## Product Designer              -0.0053053732
## Product Manager               -0.0073788694
## Recruiter                      0.0088212544
## Sales                         -0.0067363452
## Software Engineer             -0.0042460880
## Software Engineering Manager  -0.0081483472
## Solution Architect             0.0059025099
## Technical Program Manager      0.0015157547
##
## with conditional variances for "title"
```

Fixed effects of model

```
##       (Intercept) yearsofexperience    yearsatcompany            gender
##       11.56593386        0.04194513       -0.01426945        0.06556926
##              race         edu_level
##       -0.02153754        0.10465956
```

Coefficients of model

```
## $title
##                              (Intercept) yearsofexperience yearsatcompany
## Business Analyst                11.18789        0.04566153   -0.006542084
## Data Scientist                  11.45157        0.04306941   -0.017057078
## Hardware Engineer               11.22042        0.04534167   -0.016474631
## Human Resources                 11.45194        0.04306577   -0.008644633
## Management Consultant           11.23338        0.04521430   -0.007824048
## Marketing                       10.86908        0.04879557   -0.015193286
## Mechanical Engineer             10.62416        0.05120326   -0.012574920
## Product Designer                11.18299        0.04570967   -0.019574828
## Product Manager                 11.21257        0.04541892   -0.021648324
## Recruiter                       12.13249        0.03637555   -0.005448200
## Sales                           10.32850        0.05410976   -0.021005800
## Software Engineer               11.65287        0.04109051   -0.018515543
## Software Engineering Manager    13.23150        0.02557175   -0.022417802
## Solution Architect              12.82260        0.02959144   -0.008366945
## Technical Program Manager       12.88706        0.02895780   -0.012753700
##                                  gender        race  edu_level
## Business Analyst             0.05366531 -0.02153463 0.08179721
## Data Scientist               0.05596025 -0.02153837 0.15989478
## Hardware Engineer            0.04541587 -0.02153715 0.13115473
## Human Resources              0.07512946 -0.02153695 0.11540867
## Management Consultant        0.08135825 -0.02154223 0.14518842
## Marketing                    0.07626715 -0.02153629 0.12091185
## Mechanical Engineer          0.06241146 -0.02153646 0.13447478
## Product Designer             0.09100152 -0.02153784 0.08789061
## Product Manager              0.07126340 -0.02154487 0.09187752
## Recruiter                   0.06078247 -0.02153331 0.06997292
## Sales                       0.06466240 -0.02153980 0.11793981
## Software Engineer           0.04070037 -0.02154099 0.11096504
## Software Engineering Manager 0.08920465 -0.02153482 0.07656385
## Solution Architect          0.07744775 -0.02153257 0.06391408
## Technical Program Manager   0.03826855 -0.02153685 0.06193906
##
## attr(,"class")
## [1] "coef.mer"
```