IDS 702 Final Project: Data Science & STEM Salaries in the US

Peining Yang

December 2021

Summary

In this project we aim to investigate the factors influencing salaries in Data Science and STEM related jobs, taking into account of the heterogeneity of compensation across locations and companies. After fitting a hierarchical linear regression model, results show that position title, years of experience, years at current company, gender, race, and education are all associated significantly with salary, with random variations based on the state and company of employment.

Introduction

levels.fyi is a website that provides data on specific career levels and its compensations across different companies. Information on the website are submitted by individual users through uploading their offer letters, W2 Statements, Annual Compensation Statments, etc., which provides insights on an important topic that is often undiscussed. In this project, we will analyze data of Data Science and STEM salaries across companies in the United States. We will fit a hierarchical linear regression model in order to investigate the factors influencing the total annual compensation of a position, accounting for potential clustering by location and by companies.

Data

The data used for this project was scraped from the *levels.fyi* website, cleaned and uploaded to Kaggle. The original dataset contained 62,642 observations and 29 variables.

Data Wrangling

As we aim to focus solely on salaries in the United State, we first filtered out any observations that are located outside of the US, which left us with 52,838 observations. We then created a new variable of *region* which includes 5 US regions in order to perform analysis on the region level. There were also a notable amount of missing values in the dataset, concentrated between the *gender*, *race* and *education* variables. We first removed 9,464 observations that were missing all three demographics information, which leaves 27,010 observations. At this point, there are still 1,548 observations missing in *gender*, 14,960 missing in *race* and 9,869 missing in *education*. This is a significant amount of observations and we are hesitant to remove them entirely as they are likely missing not at random (MNAR). We will keep the missing data for now and perform analysis both including and excluding the problematic variables and with datasets containing missing values and with them removed. Intuitively, we expect there to be a variation of salary across different companies. Therefore, we also filtered out any companies that had less than 100 observations in order to guarantee sufficient data for analysis. After selecting only the variables of interest for our analysis, the final dataset contains 27,010 observations and 15 variables. If we remove all missing values, the dataset contains 9,551 observations, which we believe is still sufficient for analysis. A full data description can be found in Section 1.1 of the appendix.

Exploratory Data Analysis

We first explored our outcome of interest, which is the *totalyearlycompensation* variable. After plotting a histogram, we observed a severe right skew of the distribution. This prompted a log transformation on the response variable. Results

showed a significant improvement of a much more normal distribution. We will proceed with *log(compensation)* for our analysis. The figures below show the comparison before and after the transformation.

Figure 1: Distribution of Total Yearly Compensation

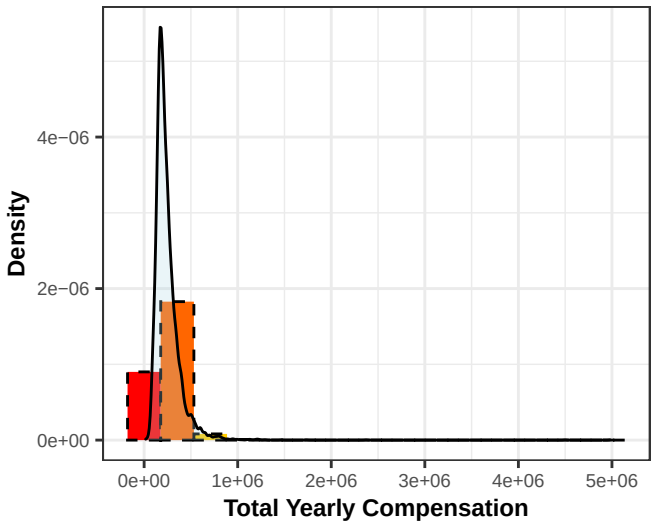
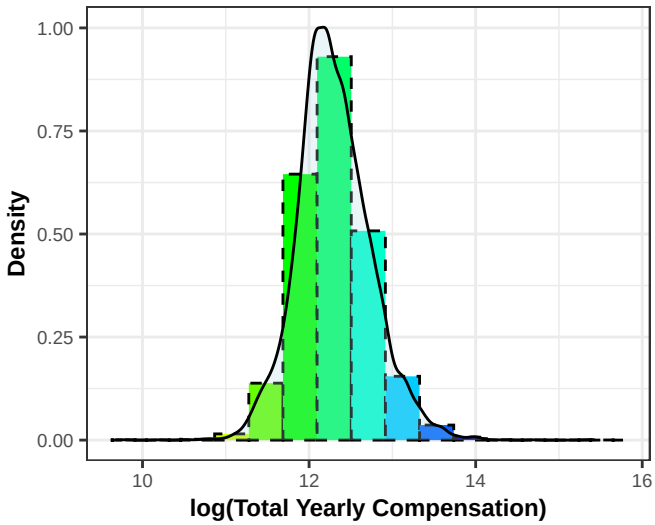
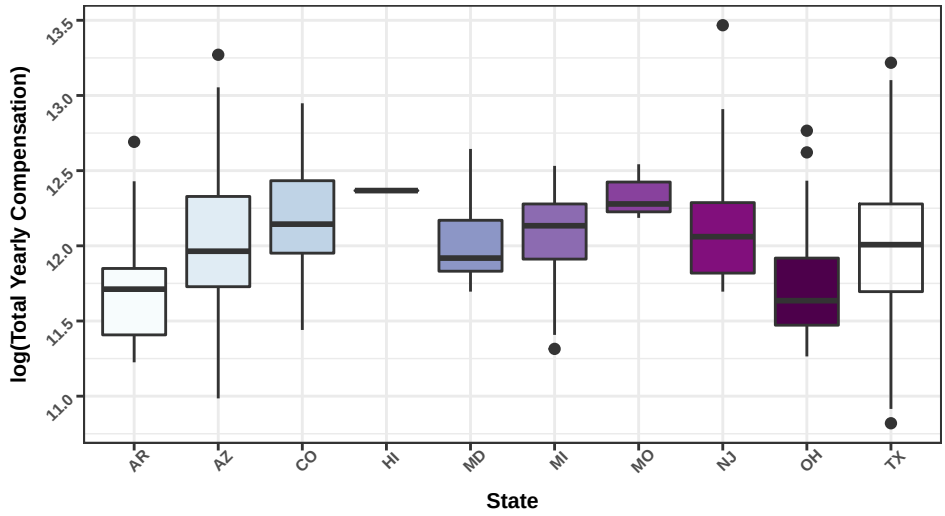


Figure 2: Distribution of log(Total Yearly Compensation)



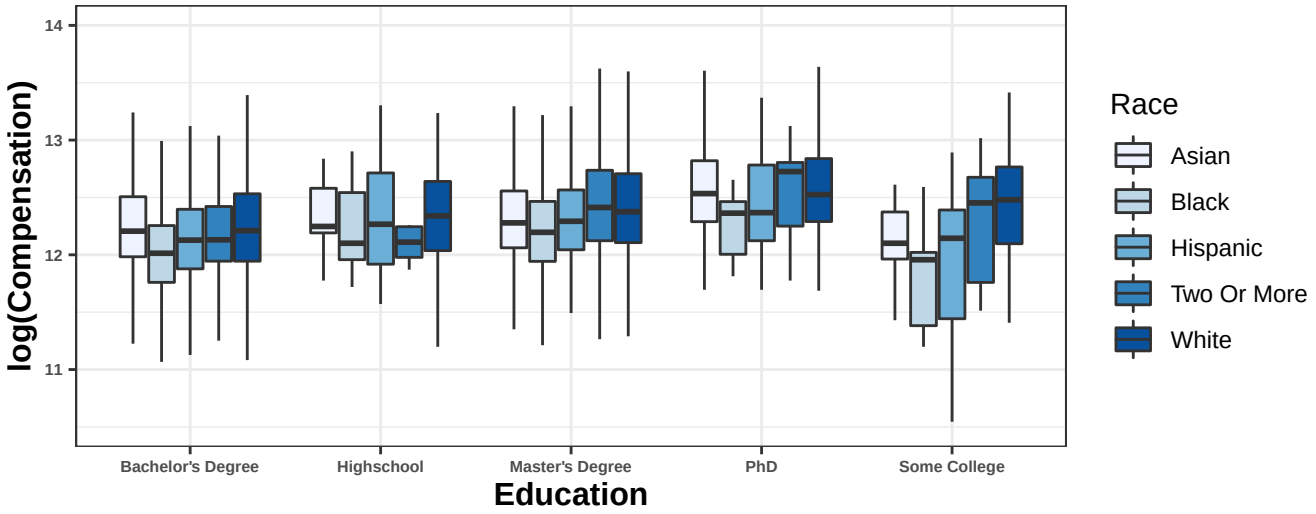
One of our research criteria is to account for variation of yearly compensations by location and/or by companies. The figure below shows the log(compensation) of a random subset of states. Results indeed showed a variation of salaries across states. Similar boxplots of log(compensation) across regions and across companies also showed variations, although less drastic for regions (See section 1.2 of the Appendix). We will include random effects by state, region, company, or more than one in our model fitting process.

Figure 3: log(Total Yearly Compensation) by State



Lastly, we explored potential interaction terms that should be included in the model. As seen from the figure below, the mean log(compensation) of employees with different degrees of education is influenced by their race. Similar variations were discovered between *race* and *gender* and *education* and *gender*. In addition, when plotting log(compensation) against years at the company, we observed a change in slopes across different position titles (See section 1.3 in Appendix). We will proceed with these interaction terms fpr the model fitting process.

Figure 4: Interaction between Education level and Race



Model

The initial step to model fitting was to determine whether removing the observations with missing values would cause a significant change in our model. We first fitted a linear regression model with *log(compensation)* as the response variable and *company*, *title*, *yearsofexperience*, and *yearsatcompany* as the predictor variables on the dataset with missing demographics information. This model generated an adjusted R-squared value of 0.619 and an AIC score of 7419.47. We then fitted two more models based on the previous model with the additional *gender*, *race* and *education* as predictor variables on both a dataset with missing values and a dataset with missing values removed. Using the full dataset, the model generated an adjusted R-squared value of 0.646 and an AIC score of 2976.63. Using the dataset with missing values removed, the model generated an adjusted R-squared value of 0.617 and an AIC score of 2485.68. In addition, the latter two models only had a 6.67% difference in model coefficients. Given the low AIC score, we decided to continue our analysis using the dataset with all missing values removed.

From the exploratory data analysis, we tested the interaction terms of *title:yearsatcompany*, *gender:education*, *race:education* and *gender:race*. A stepwise AIC model selection algorithm chose to keep all four interaction terms in the model. However, the interaction terms showed extremely high Variance Inflation Factors (VIF), indicating that there is significant issues with multicollinearity (See section 2.1 of Appendix for VIF values). Removal of any interaction terms did not help with lowering VIF. Therefore, we have chosen to remove interaction terms from the model.

Hierarchical Linear Regression Model

Since we observed varying means of log(compensation) across states, US regions and companies in the EDA, we first wanted to determine whether it's statistically significant to include varying intercepts for *state*, *region*, or *company* in the model. Building on top of our base linear regression model, we first fitted a model with only *state* as the random effect and another model with both *state* and *region* as the random effects. Both models had similar AIC values of 2313.53 and 2313.99, respectively. An ANOVA test between the two showed a p-value of 0.317, which is above the 0.05 threshold. We can conclude that the inclusion of *region* as a random effect is not statistically significant. We then fitted a third model with *state* and *company* as varying intercepts. An ANOVA test produced a p-value of less than 0.001, indicating that the random effect of company is statistically significant. In additon, this model has an AIC value of 2266.94, which is smaller than the previous two models. We will proceed with *state* and *company* as random effects in the model.

Next we wanted to determine whether we should include any varying slopes. From the EDA, we discovered that there is some variation of log(compensation) for different titles across states. After fitting a model with varying slopes of *title* and *state*, an ANOVA test with the previous model showed a p-value of 0.748, which is above the 0.05 threshold and statistically insignificant. This model also showed an AIC value of 2645.59, which is higher than the model with varying intercepts (AIC of 2313.53). Therefore, we will exclude varying slopes from the model. The final model contains *title*, *yearsofexperience*, *yearsatcompany*, *gender*, *race*, and *education* as predictor variables with random effects of *state* and *company*. The mathematical notation of this model is shown below. A full model output can be found in section 2.2 of the Appendix.

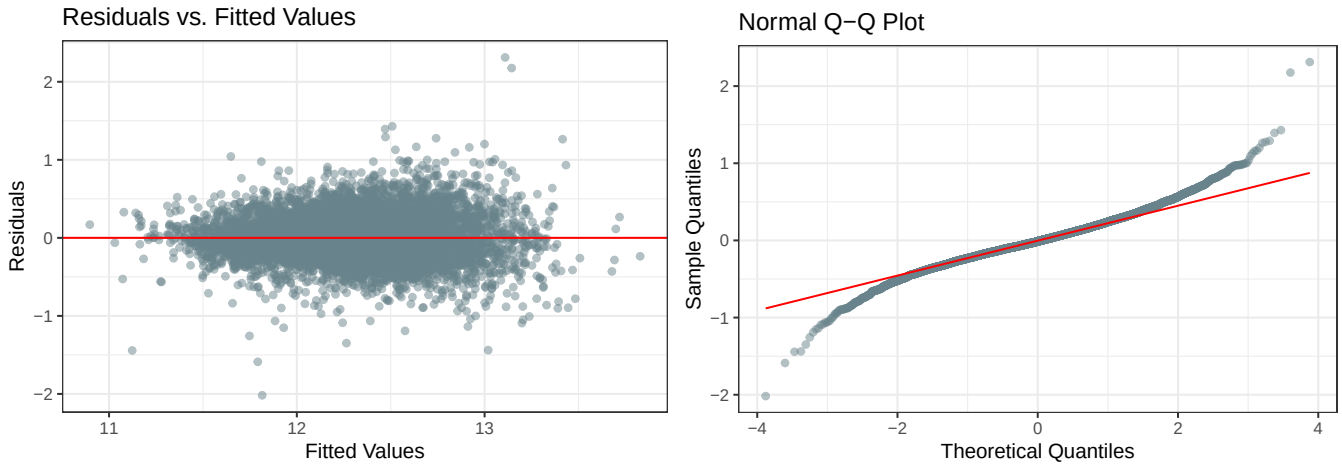
$$log(Compensation_{ijk}) = (\beta_0 + \gamma_{0k} + \gamma_{0jk}) + \sum_{a=2}^{15} \beta_{1a}[title_i = a] + \beta_2 yearsofexperience_i + \beta_3 yearsatcompany_i +$$
$$\sum_{b=2}^{3} \beta_{4b}[gender_i = b] + \sum_{c=2}^{5} \beta_{5c}[race_i = c] + \sum_{d=2}^{5} \beta_{5d}[education_i = d] + \epsilon_i jk; i = 1, ..., n_j; j = 1, ..., J$$

$$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$
$$(\gamma_{0k}, \gamma_{0jk}) \sim \mathcal{N}_{\in}(\mathbf{0}, \Sigma)$$

where a takes on different levels of the title variable, b takes on different levels of the gender variable, c takes on different levels of the race variable and d takes on different levels of the education variable.

Model Assumptions

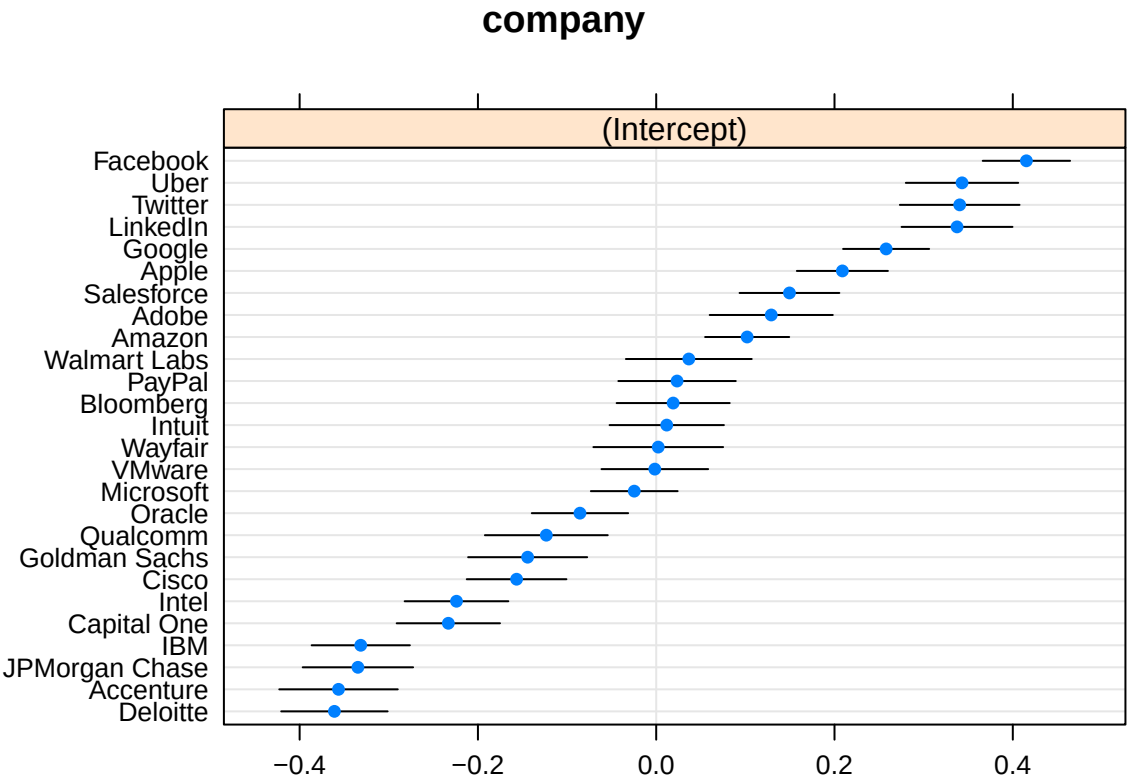
With our final model, we performed various diagnostic tests to assess model assumptions. From the residuals vs. fitted values plots below, we can see that the constant variance and independence assumptions are not violated as points are scattered relatively randomly along the horizontal line at 0. The QQ plot of residuals of the final model showed that the points on both ends are slightly trailing away from the 45 degree line. This indicates that there is a violation of the normality assumption with the model. We have already performed a log transformation on the response variable, which now shows a normal distribution. We are hesitant to classify any points as outliers due to the nature of salaries. Therefore, we will keep our model with only a log transformation. The VIF values of the final model are all below 3, indicating that we do not have issues of multicollinearity.



Model Interpretation

The baseline of the model is someone who is an Asian female Data Scientist with a Bachelor's degree and 0 years of experience and 0 years at the current company. We expect this employee's total yearly compensation to be $e^{11.675} = \$117,594.8$. Keeping all else constant, with every unit increase in experience, we expect her compensation to increase by a multiplicative effect of $e^{0.0334} = 1.034$, which is about a 3.4% increase. With every unit increase in years at the company, we expect a multiplicative effect of $e^{0.0066} = 1.0066$, which is about a 0.6% increase. If her position is a Software Engineer and keeping all else constant, we expect her total yearly compensation to increase by a multiplicative effect of $e^{0.0498} = 1.051$, which is about a 5.1% increase. Keeping all else constant, we expect roughly a 6.7% increase if this person was male, a 4.3% decrease if this person was Black and a 26.6% increase if this person has a PhD.

The estimated standard error for state is 0.149, which describes the across state variation attributed to the random intercept. For companies, the estimated standard error is 0.232. This implies that the total yearly compensation of an employee in the tech industries varies more by company than by state. The estimated standard error of the residual of the model is 0.266, which describes the within-state/company or the remaining unexplained variation. As seen from the figure below, our model is statistically significant for all besides 7 companies. A similar plot in the section 3.1 of the Appendix shows the states where our model is statistically significant.



Limitations

One of the biggest limitations to this project is the amount of missing values in the original dataset. On average, there were about 50% of data missing for the demographics variables. Since these values are likely missing not at random, we are hesitant to perform imputations. Although removing these observations did not pose a significant effect on the model, it would be more ideal to include them as removing them entirely could've taken out crucial information from the other predictor variables. In addition, we did not include any interaction terms in the final model due to issues of multicollinearity.

Conclusion

According to the final model, an employee expected to earn the most salary is a White male with a PhD working as a Software Engineering Manager. With every year increase in experience and employment at current company their salary is also expected to increase. This project generated practical insights that could potentially benefit job seekers in the Data Science and STEM field. In the future, it would be interesting to include international data or price of living in each location of employment for a more holistic analysis.

Appendix

1.1

Table 1: Data Description

Variable	Description
company	Company of employment (character)
level	Position level within company.
title	Position title within company,
totalyearlycompensation	Total yearly compensation in US$ (outcome)
location	City and state of employment.
yearsofexperience	Total years of working experience.
yearsatcompany	Total years at current company.
basesalary	Baseline salary in US$.
stockgrantvalue	Stock grant value in US$.
bonus	Bonus in US$
gender	Gender: Male, Female, or Other
race	Race: White, Black, Asian, Hispanic, or Two or More
education	Education level: Highschool, Some College, Bachelor's Degree, Master's Degree, or PhD
state	State of employment.
region	US region of employment.

1.2

Figure 4: log(Total Yearly Compensation) by Company

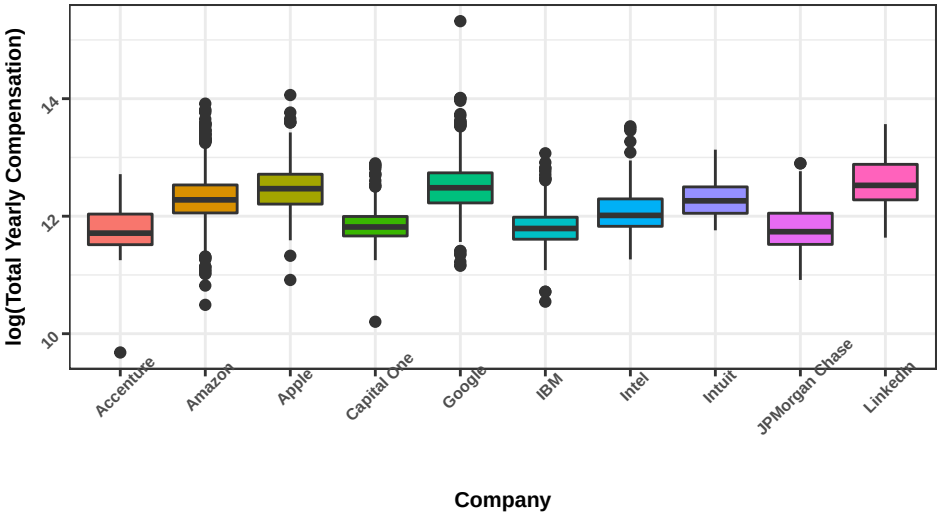
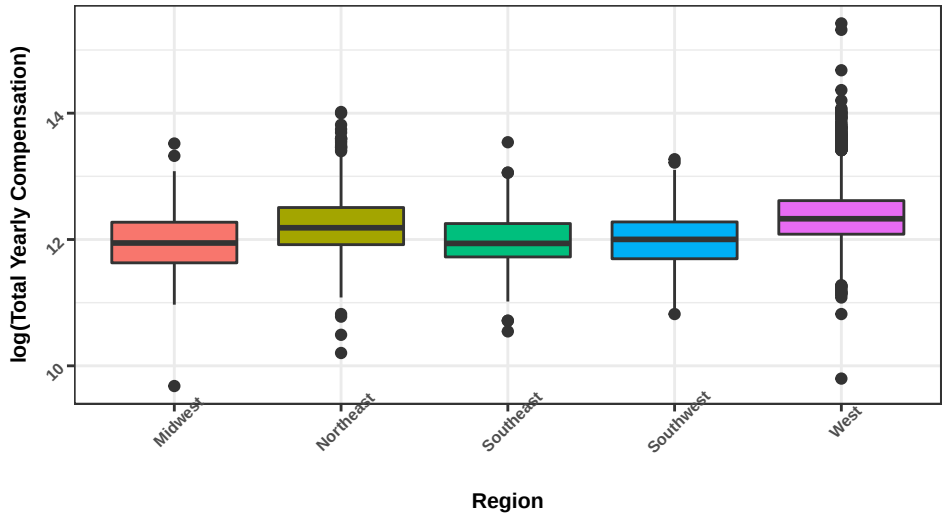
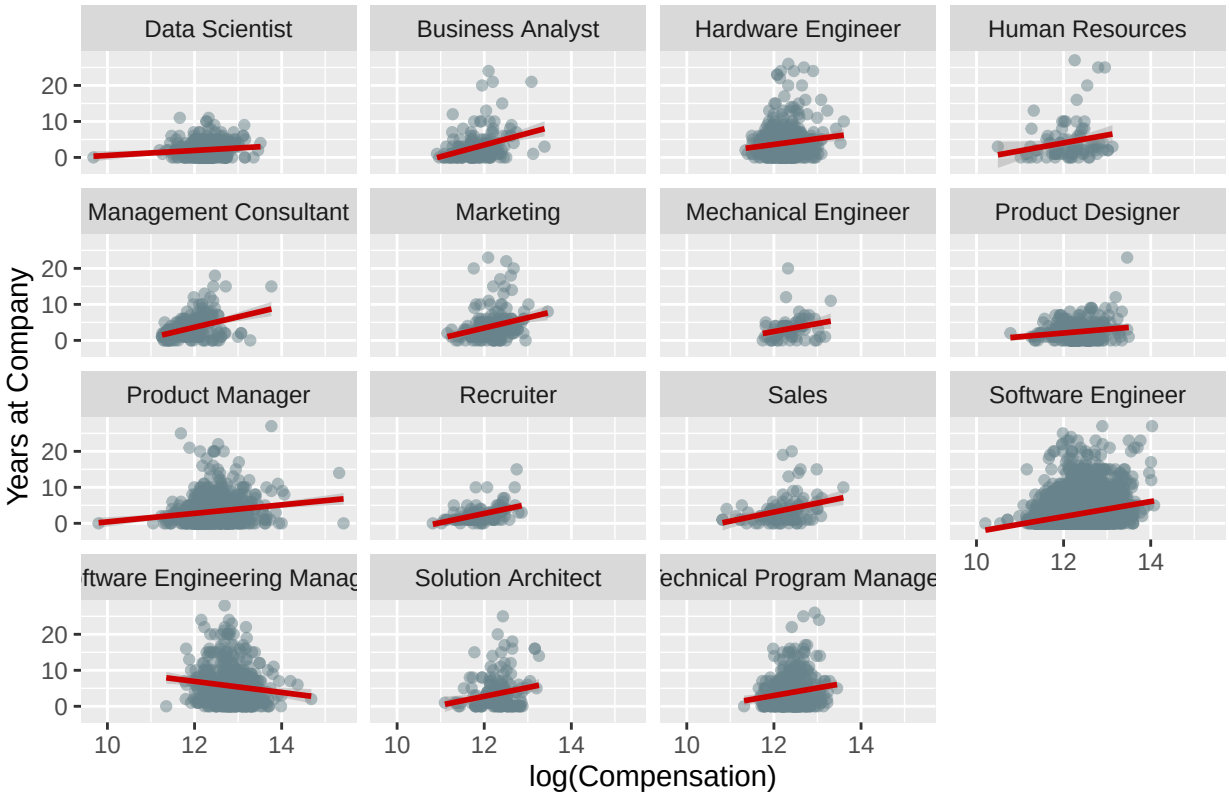


Figure 4: log(Total Yearly Compensation) by Region



1.3

Figure 2.2: log(Compensation) vs. Years at Company by Position Titles



2.1

Table 2: VIF of Model with Interactions

	GVIF	Df	GVIF^(1/(2*Df))
company	3.850	25	1.027
title	12990.534	14	1.403
yearsofexperience	1.813	1	1.346
yearsatcompany	79.958	1	8.942
gender	149.040	2	3.494
Race	1204.003	4	2.427
Education	58737.226	4	3.946
title:yearsatcompany	256296.354	14	1.560
gender:Education	111859.727	8	2.068
gender:Race	63401.800	8	1.996
Race:Education	3624.619	16	1.292

2.2

Table 3: Hierarchical Linear Regression Model Output

Fixed Effects	Estimate	Std. Error	t value	2.5 %	97.5 %
(Intercept)	11.675	0.055	211.157	11.566	11.785
titleBusiness Analyst	-0.303	0.025	-12.329	-0.351	-0.255
titleHardware Engineer	-0.035	0.021	-1.686	-0.075	0.006
titleHuman Resources	-0.307	0.032	-9.540	-0.370	-0.244
titleManagement Consultant	0.056	0.029	1.911	-0.002	0.114
titleMarketing	-0.182	0.024	-7.448	-0.230	-0.134
titleMechanical Engineer	-0.065	0.035	-1.851	-0.134	0.004
titleProduct Designer	0.012	0.021	0.593	-0.029	0.054
titleProduct Manager	0.116	0.017	6.613	0.081	0.150
titleRecruiter	-0.411	0.030	-13.695	-0.469	-0.352
titleSales	-0.022	0.031	-0.727	-0.083	0.038
titleSoftware Engineer	0.050	0.015	3.378	0.021	0.079
titleSoftware Engineering Manager	0.285	0.019	14.720	0.247	0.323
titleSolution Architect	0.000	0.025	0.004	-0.049	0.049
titleTechnical Program Manager	-0.056	0.020	-2.822	-0.095	-0.017
yearsofexperience	0.033	0.001	58.135	0.032	0.035
yearsatcompany	0.007	0.001	6.912	0.005	0.008
genderMale	0.064	0.007	9.002	0.050	0.078
genderOther	0.165	0.047	3.521	0.073	0.256
RaceBlack	-0.044	0.015	-2.896	-0.073	-0.014
RaceHispanic	-0.056	0.012	-4.549	-0.080	-0.032
RaceTwo Or More	0.000	0.015	-0.005	-0.030	0.030
RaceWhite	0.005	0.007	0.786	-0.008	0.018
EducationHighschool	-0.027	0.027	-1.005	-0.081	0.026
EducationMaster's Degree	0.058	0.006	9.412	0.046	0.070
EducationPhD	0.236	0.013	18.590	0.211	0.261
EducationSome College	-0.139	0.024	-5.807	-0.186	-0.092

Random Effects Groups	Variance	Std.Dev
state	0.022	0.149
company	0.054	0.232
Residual	0.071	0.266

3.1