MA678 Final Project

Yin Xu

2022-11-30

Abstract

In most universities, AI, Statistics and Data Science are some of the most popular majors, plenty of students choose these majors since if they have the Data Science background, they may find jobs easily after university. According to this issue, there is a dataset from data science jobs shows that the titles, locations and other variables which effect the salaries. To discuss this topic, I build a multilevel model with group `title` and `race` The report includes 4 main parts: introduction, method, result and discussion.

Introduction

As we always know, the job titles, locations of companies, experience level and other factors usually influence the amount of a person's salary. However, after COVID-19 came into everyone's life, some of jobs became remote to avoid direct contact, therefore, remote jobs may become one factor which influence the salaries. Remote work may decrease the salaries, because the efficiency will be reduced; sometimes, because of the extension of the work hour, the salaries will be increased. Besides, some small companies prefer to attract talents with high wages; at the same time, large companies have complete talents and do not care about the whereabouts of employees with low salaries.

Therefore, multilevel model is needed to discover the effect of the factors. As the assumption, I considered that because of the diverse titles, experience may influence much differently; and the gender and race may cause the same the same results. To figure this out, I divide the factors to two group: fixed effect factors(e.g. education, bonus, location) and random effect factors(title, race).

Method

Data Cleaning

The data is from kaggle open dataset(https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries).

This data includes the variables of salaries from 2020 to 2021 around the world.

First of all, there are "NA"s in the data, cleaning data is essential. Then I want to focus on the jobs in the United States, so the observations outside of USA should be dropped. Besides, I dropped several observations with super high wages (`total yearly compensation`) which is not fit for the plots.

Secondly, there are 15 unique titles in `title`, but I only want to keep the position of the titles (e.g. manager, engineer, sales and so on). Therefore, I extracted the last word from `title` and added a new column named `title_1`, this column will be used as title for further analysis.

Additionally, in this data, `gender` and `Education` are two variables in characters, to make my exploratory data analysis more reasonable and easy to read, I changed `gender` to numeric values 0, 1, 2 which means other, female and male; for `Education`, 0 means "High School", 1 means "Some College", 2 means "Bachelor's Degree", 3 means "Master's Degree", and 4 means "PhD".

After data cleaning, the following is the chart of variables which are used in this report:

Variables Names	Explanation
`timestamp`	When the data was recorded.
`company`	Company names.
`title`	Job title.
`location`	Job location.
`cityid`	City ID of the location.
`totalyearlycompensation`	Total yearly compensation which is added by `basesalary`, `stockgrantvalue`and `bonus`.
`basesalary`	Base salary of the year.
`stockgrantvalue`	Stock grant value.
`bonus`	Bonus of the year.
`yearsofexperience`	Year of experience of Data Science jobs.
`yearsatcompany`	Year of experience at said company.
`gender`	Gender of the observations.
`race`	Race of the observations.
`Education`	Education background.

Exploratory Data Analysis

After data cleaning, I got 16937 observations and 14 variables, and I set `total yearly compensation` as the output of the analysis; the other 13 variables will be discussed in following analysis to know if they influence the salaries of these Data Science jobs. To figure out the relationships between the variables and build a model, I select `title` and `race` as two groups.

Figure 1: Salary vs Years at Company

These two graphs show the relationship between `total yearly compensation` and `years at company` which is varied by `title` and `race`. Both of the graphs show that as the years at the company increasing, the salaries increase, and the slopes are lightly different. By the first graph, the interprets of the different titles are different, which means that the very first salary after getting the Data Science jobs depends on the title, but not race(b). Besides, when I use `year of experience`, `location`, `timestamp`, `stock grant value` and `bonus` to compare the relationships with `total yearly compensation`, they showed the similar trend.

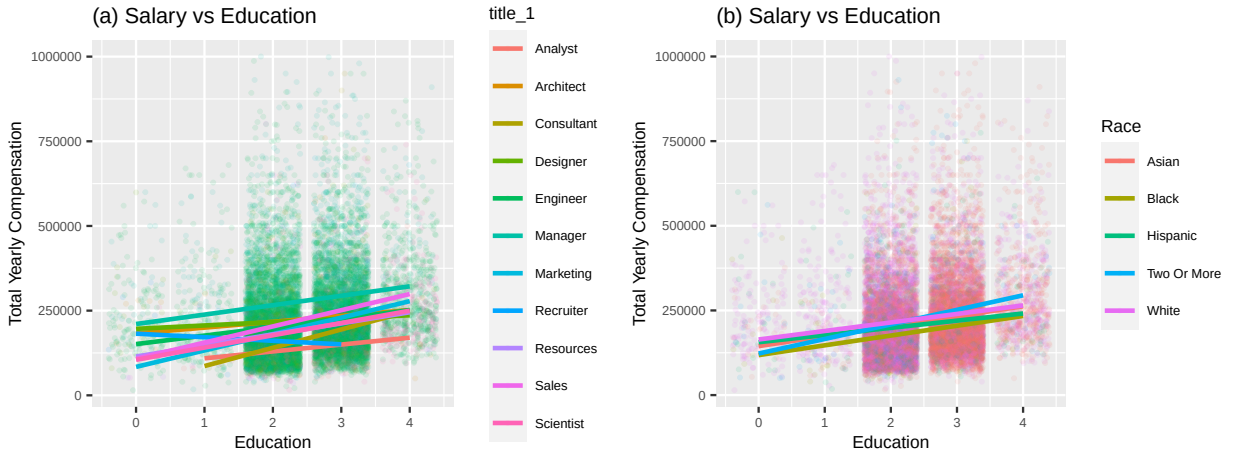


Figure 2: Salary vs Education

According to the salary and education graphs, as the education background increasing, the slopes of most titles are increasing, there only "Recruiter" line shows the different trend. Additionally, "Analyst" and "Consultant" have no High School background, and most engineers have bachelor or master background. Overall, `title` influences `Education`, however, there is no difference of education background among diverse races.

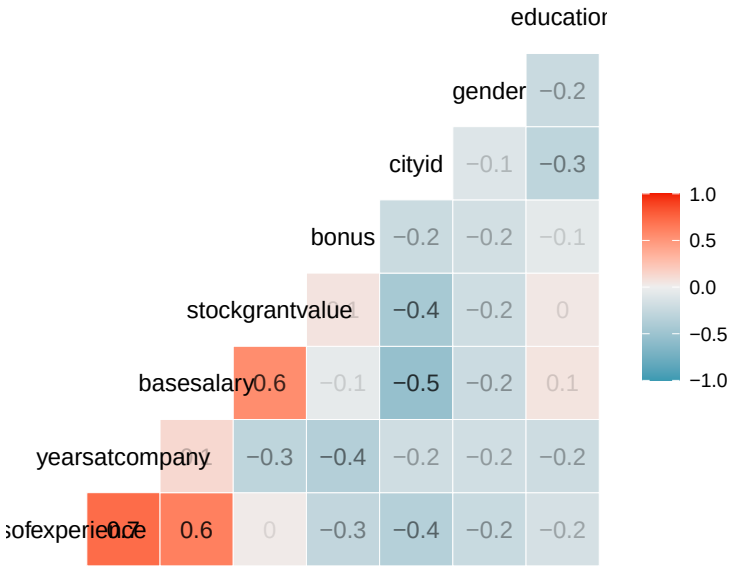
Figure 3: Salary vs Gender

In Figure 3, we can know that the relationship between `total yearly compensation` and `gender` does not depend on titles, because of the similar trends of the plot; however, "Two or More" race has the decreasing slope which is different from other races. Therefore, I would like to say that the `race` types have affect on the relationship between `total yearly compensation` and `gender`.

Correlation Checking

Because of the large number of the `total yearly compensation` and the distributions of the variables are somehow skewed. To fit the multilevel model, I would like to change the variable to `log` as the new variables, for example, `log(education + 1)`, and build a new data frame.

By further consideration, there may be some variables have correlation except `title` and `race`; if there is a strong correlation between two variables, just one of them should be put into the final model. The following graph shows the number of variables' correlation:



According to this graph, `years at company` and `years of experience` have the correlation at 0.7, which is considered as the largest number in this graph; also, 0.7 is close to 1. Besides, `base salary` & `year`

`of experience` and `base salary` & `stock grant value` also show the large correlation at 0.6. As the result, these three variables are considered that they have the mutual influence, and I will keep `years of experience` and `stock grant value` in the multilevel model.

Multilevel Model

Based on the analysis above, I built a multilevel model after fitting:

```
model_1 <- lmer(salary ~ yearsofexperience +stockgrantvalue + cityid + education + gender
              + (yearsofexperience + stockgrantvalue + cityid + education | title)
              + (1 | race),
              data = cor, REML = F)
```

The following is the summary of the model of fixed effects, the table shows that the probabilities are small enough to say that these variables are statistically significant at $\alpha = 0.05$.

	Estimate	Std. Error	df	t value	Pr(>
(Intercept)	1.143e+01	9.984e-02	6.404e+00	114.484	7.57e-12 ***
yearsofexperience	3.036e-01	2.489e-02	9.900e+00	12.195	2.76e-07 ***
stockgrantvalue	4.784e-02	3.150e-03	1.068e+01	15.186	1.43e-08 ***
cityid	-5.774e-02	9.895e-03	4.476e+00	-5.835	0.003010 **
education	6.336e-02	1.379e-02	1.009e+01	4.596	0.000963 ***
gender	7.190e-02	1.371e-02	1.683e+04	5.244	1.59e-07 ***

Model checking

After the model building, model checking is needed. I checked the multilevel by Q-Q plot, it shows that the mean of the standardized residuals is around 0 which considered as a normal distribution without skewed, and there is no extreme value, so this model is fitted well.

Additionally, I used `ranef` to extract the conditional modes of the random effects(`title` and `race`). I move the completed table to the appendix.

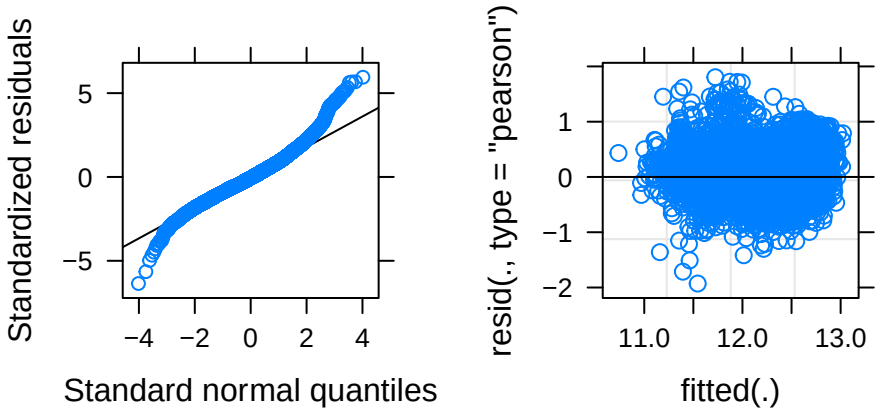


Figure 4: Correlation Graph

Result

As the conclusion of the report, I built a multilevel model with two random effects(`title` and `race`):

$$log(TotalCompensation+1) = 11.43 + 0.303 \times log(Experience+1) + 0.048 \times log(StockValue+1) - 0.058\times$$

$$log(Location+1) + 0.063 \times log(Education+1) + 0.072 \times log(Gender+1)$$

In this model, there is only `cityid` has the negative correlation with `total yearly compensation`. According to the ranef graph above, I would like to check the details of the slope of the variables in the model. This is the model I used `Engineer` and `Asian` which has the largest observations in the data:

$$log(TotalCompensation+1) = 11.66 + 0.29 \times log(Experience+1) + 0.058 \times log(StockValue+1) - 0.052\times$$

$$log(Location+1) + 0.043 \times log(Education+1) + 0.072 \times log(Gender+1)$$

Compared with the original model, there is little change on the number of slopes, which means that asian engineers model has no dramatic different; when the `years of experience` increases 1 year, the log of `total yearly compensation` will increase by 0.29. Also, according to the `ranef` table, experience of titles of `Marketing`, `Recruiter`, `Resources` and `Sales` may influence more on the salary; the table shows that `Black` and `Hispanic` have the negative intercept, therefore, they may have less increase on `total yearly compensation` than others.

Discussion

By the above multilevel model, we can know the relationships between variables and the total salaries, I choose `title` and `race` as two groups to fix other variables in the model. `Titles` and `races` do increase or decrease the change of one unit change on variables; for example, most data science engineers work at west coast states(e.g. CA, WA and so on) because of the location of silicon valley, but sales is distributed throughout the country. After the whole analysis of the report, to a certain extent, the result is different from my assumption. The `gender` which I considered as a variable that has large impact has a large probability by t-test in model summary, so it is not statistically significant.

```
yr_2 <- subset(cor, cor$yearsofexperience < 0.7)
pre_yr2 <- predict(model_1, newdata = yr_2)
exp(mean(pre_yr2))
```

```
## [1] 130617.6
```

Besides, to predict the data science first job yearly salary, I subset the people who have less than 2 years job experience, and made a prediction. The result shows their average total yearly compensation is around 130,617.6 USD, and this can provide reference for data science and statistics students who are about to graduate.

Appendix

```
## $title
##            (Intercept) yearsofexperience stockgrantvalue        cityid
## Analyst     0.18019026      -0.043019718   -0.0086968478 -0.0173509387
## Architect   0.04357079      -0.062098692   -0.0058792925  0.0224405240
## Consultant -0.04424508      -0.001882186   -0.0204956704  0.0041037475
## Designer    0.09973842      -0.030964534    0.0117010620 -0.0007664164
## Engineer    0.23361522      -0.097904222    0.0090551467  0.0062905239
## Manager    -0.09340627      -0.009219868    0.0097046154  0.0275026316
## Marketing  -0.09074738       0.069555590    0.0049567843 -0.0168201012
## Recruiter  -0.09202093       0.064755202   -0.0063805812 -0.0174019117
## Resources  -0.16925357       0.085279743    0.0043324049 -0.0089731265
## Sales      -0.46922518       0.131552860   -0.0007352052  0.0276189553
## Scientist   0.40178372      -0.106054174    0.0024375836 -0.0266438876
##               education
## Analyst    -0.011662969
## Architect  -0.002115131
## Consultant  0.072383914
## Designer   -0.008132791
## Engineer   -0.020106352
## Manager    -0.025472901
## Marketing  -0.009478295
## Recruiter  -0.018097963
## Resources  -0.041058796
## Sales       0.004753285
## Scientist   0.058987997
##
## $race
##               (Intercept)
## Asian        0.0009249878
## Black       -0.0243825727
## Hispanic    -0.0113589059
## Two Or More  0.0293894621
## White        0.0054270287
##
## with conditional variances for "title" "race"
```

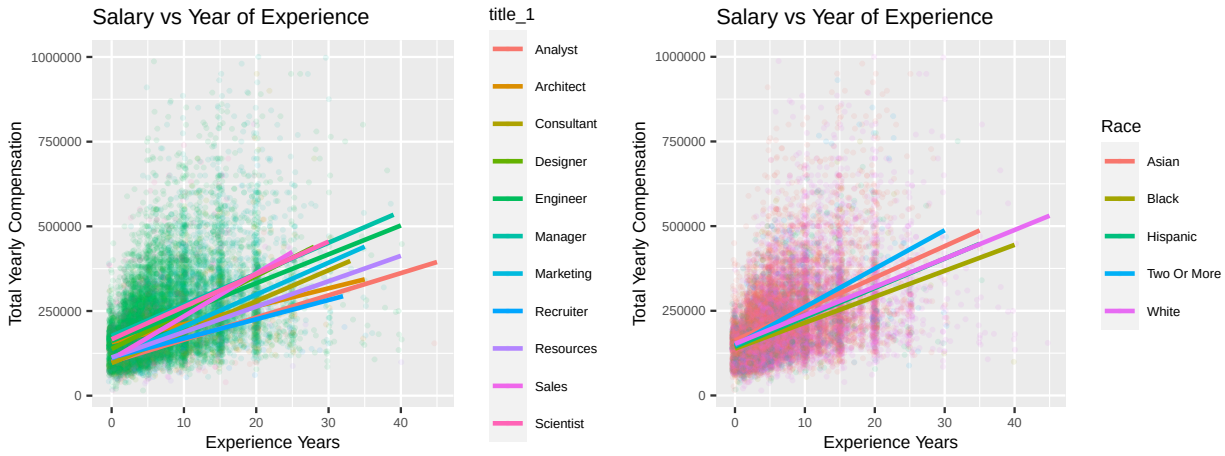


Figure 5: Salary vs Experience Years

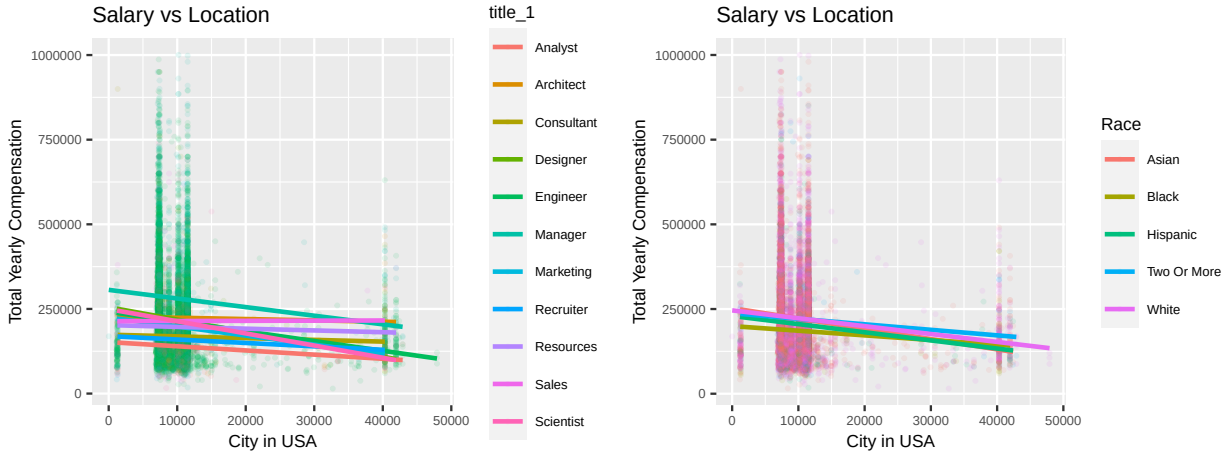


Figure 6: Salary vs Location

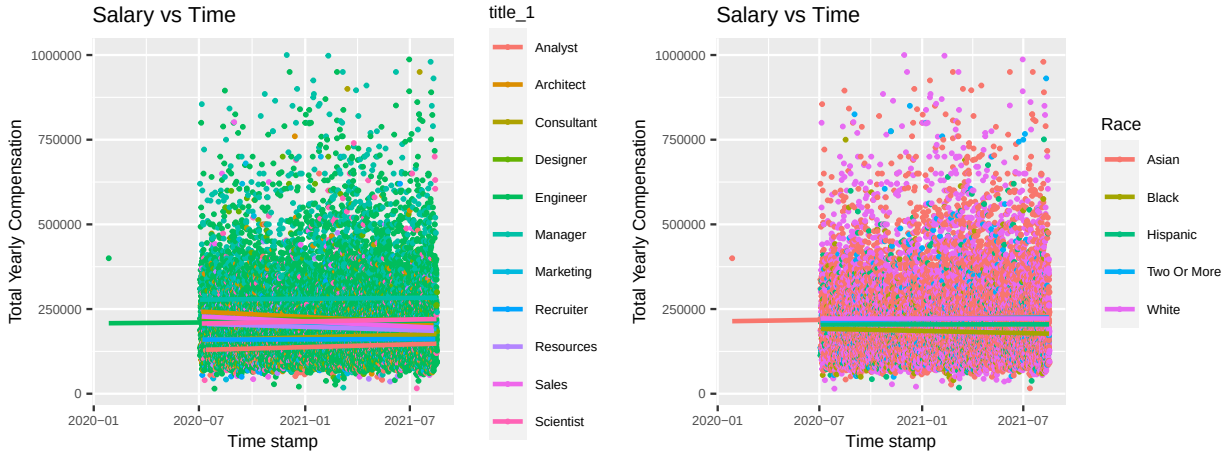


Figure 7: Salary vs Time

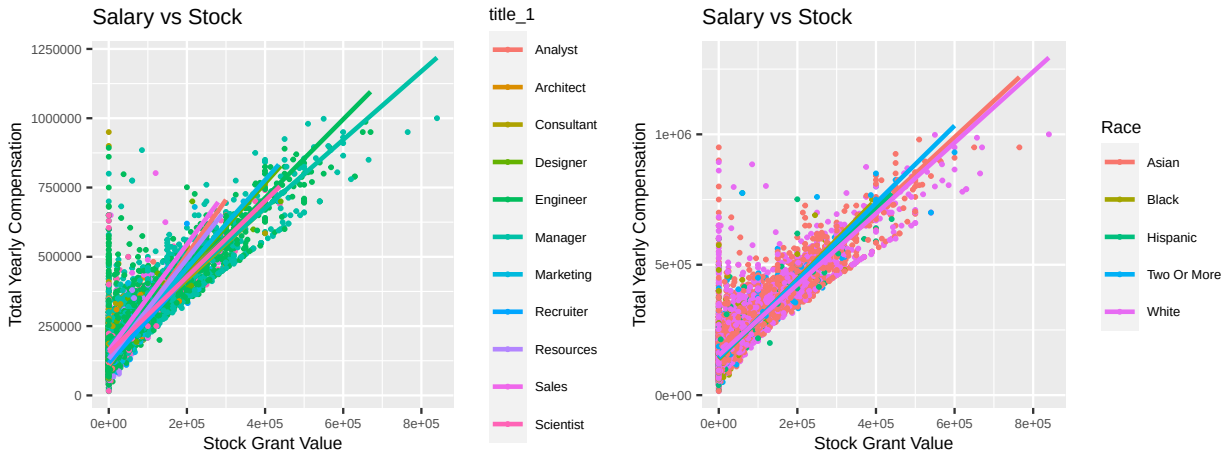


Figure 8: Salary vs Stock

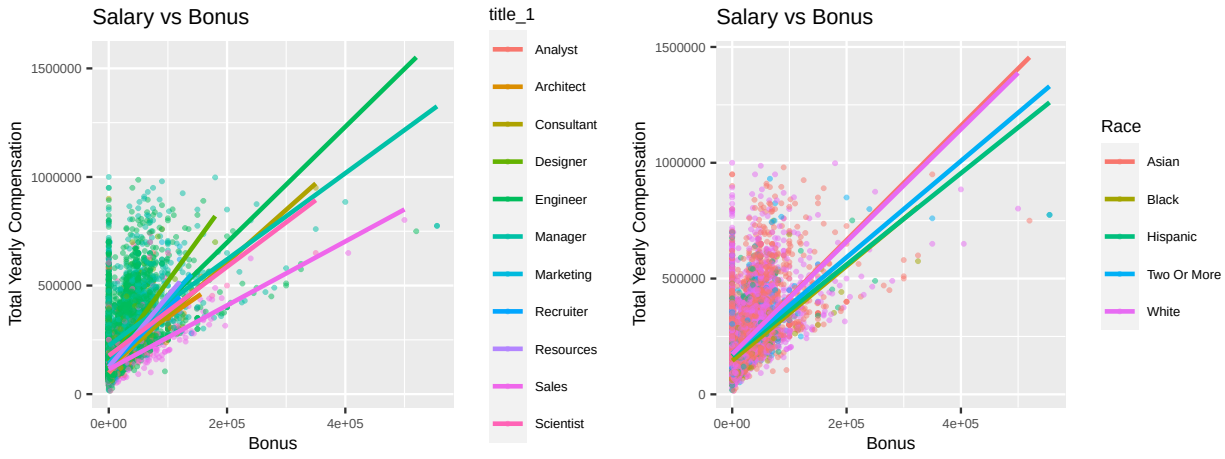
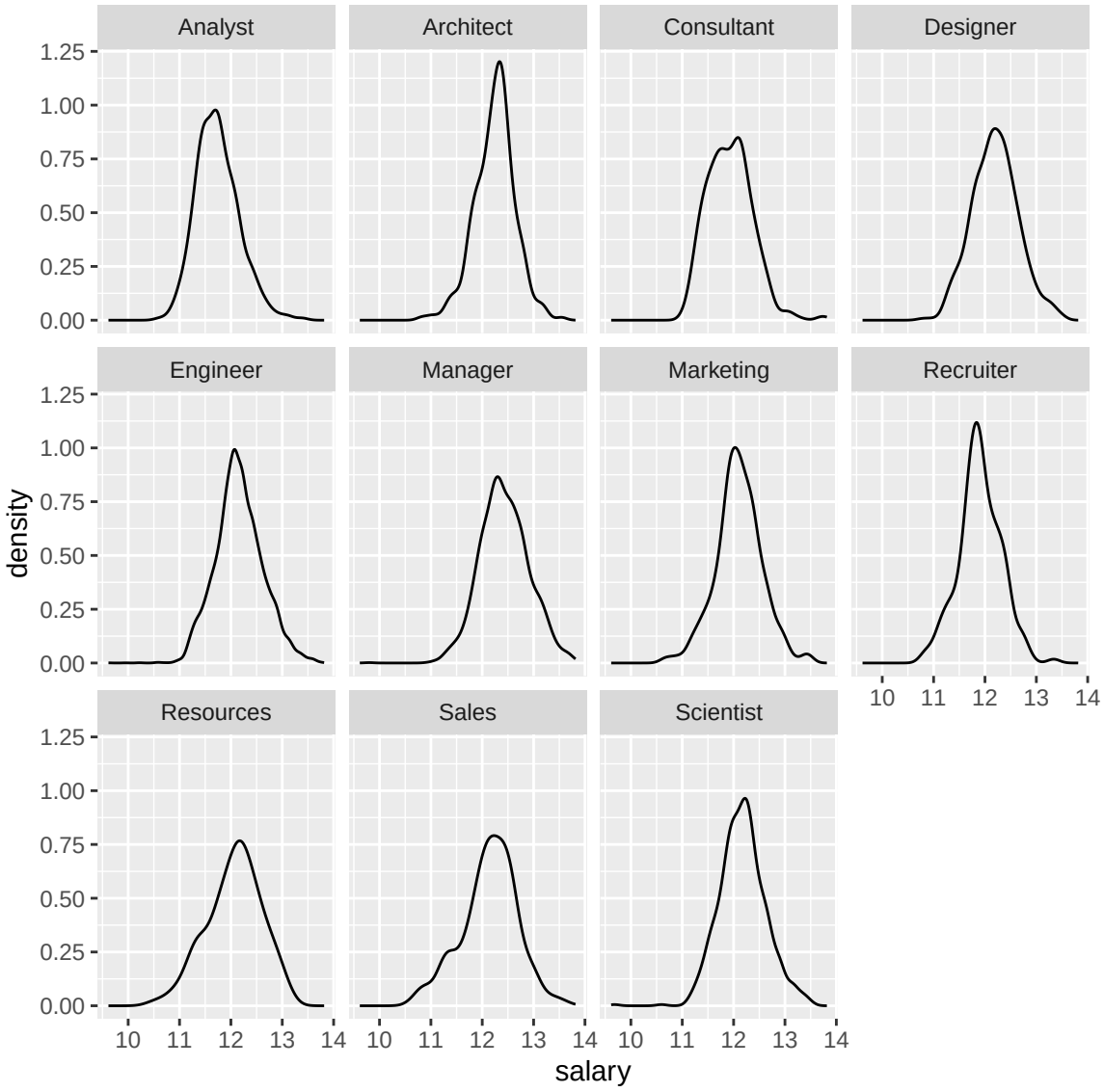
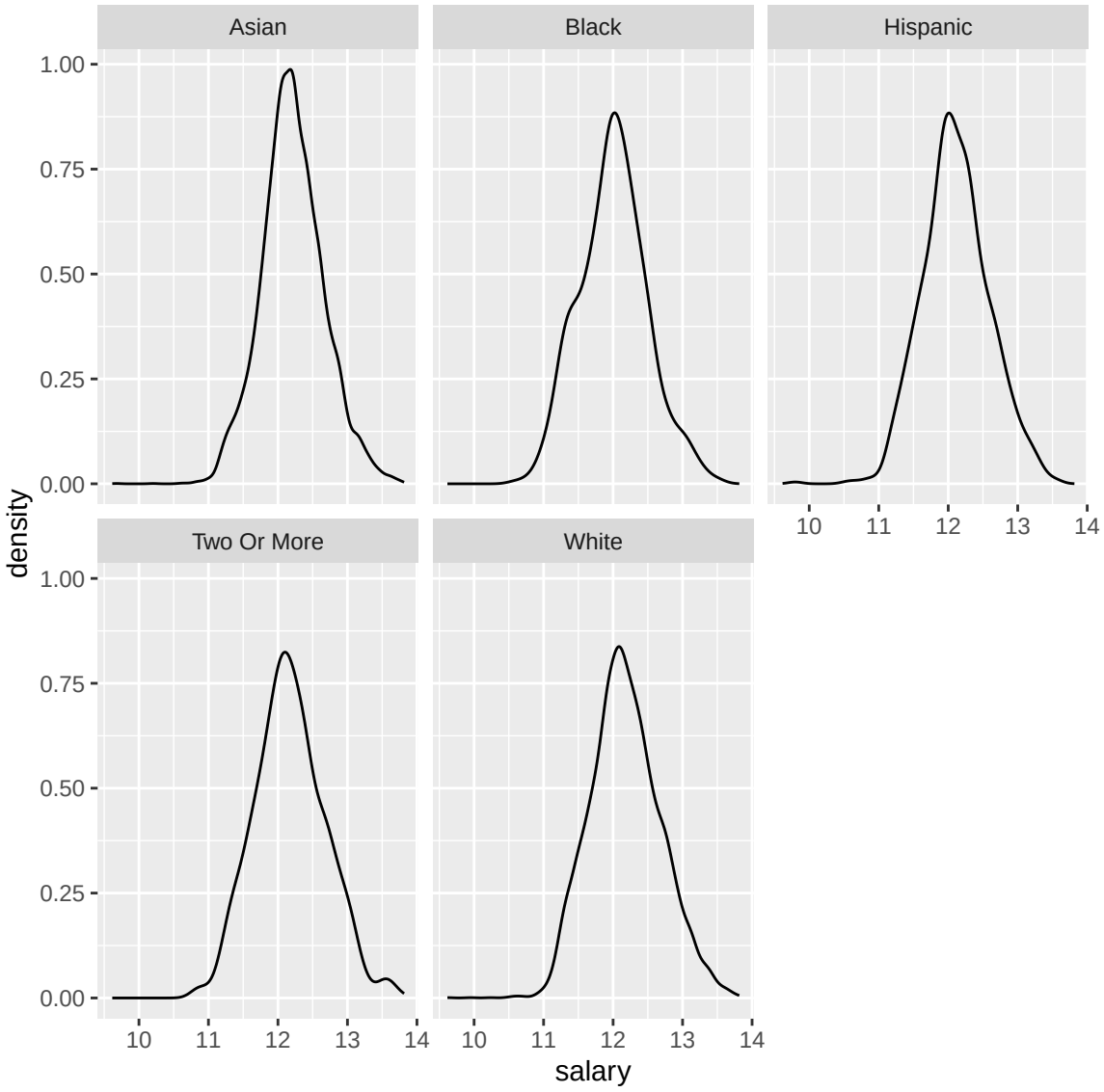


Figure 9: Salary vs Bonus


Analyst
Architect
Consultant
Designer
Engineer
Manager
Marketing
Recruiter
Resources
Sales
Scientist
density
salary
0.00
0.25
0.50
0.75
1.00
1.25
10
11
12
13
14


Asian
Black
Hispanic
Two Or More
White
density
1.00
0.75
0.50
0.25
0.00
10
11
12
13
14
salary