678_midterm

Yalong Wang

2022-12-03

Abstract

Engineering is the use of scientific principles to design and build machines, structures, and other objects, including bridges, tunnels, roads, and vehicles The discipline of engineering includes a wide range of more specialized engineering fields, each with a more specific emphasis on particular areas of applied mathematics, applied science, and types of applications. Engineering is a broad discipline that is often subdivided into several sub-disciplines. Thus, here comes the problem: which kind of graduate students can earn most salary. To figure out this problem, I built a multilevel model with group level specialization. This report will be written by 5 main parts: Abstract, Introduction, Method, Result, Discussion.

Introduction

A relevant question is what determines the salary and the jobs these engineers are offered right after graduation. Various factors such as college grades, candidate skills, the proximity of the college to industrial hubs, the specialization one have, market conditions for specific industries determine this. On the basis of these various factors, my goal is to determine the salary of an engineering graduate in India and the predcit the salary with those variables.

Methods

Data Preprocessing

I found the data set from a public website(https://www.kaggle.com/datasets/manishkc06/engineering-graduate-salary-prediction (https://www.kaggle.com/datasets/manishkc06/engineering-graduate-salary-prediction)). The interpretation of the data is below:

column names	explanation
ID	A unique ID to identify a candidate
Salary	Annual CTC offered to the candidate (in INR)
Gender	Candidate's gender
DOB	Date of birth of the candidate
CollegeID	Unique ID identifying the university/college
CollegeTier	Each college has been annotated as 1 or 2.
Degree	Degree obtained/pursued by the candidate
Specialization	Specialization pursued by the candidate
CollegeGPA	Aggregate GPA at graduation
CollegeCityID	A unique ID to identify the city in which the college is located in.
CollegeCityTier	The tier of the city in which the college is located in.
CollegeState	Name of the state in which the college is located
GraduationYear	Year of graduation (Bachelor's degree)
English	Scores in AMCAT English section
Logical	Score in AMCAT Logical ability section
Quant	Score in AMCAT's Quantitative ability section
Domain	Scores in AMCAT's domain module

Data cleaning

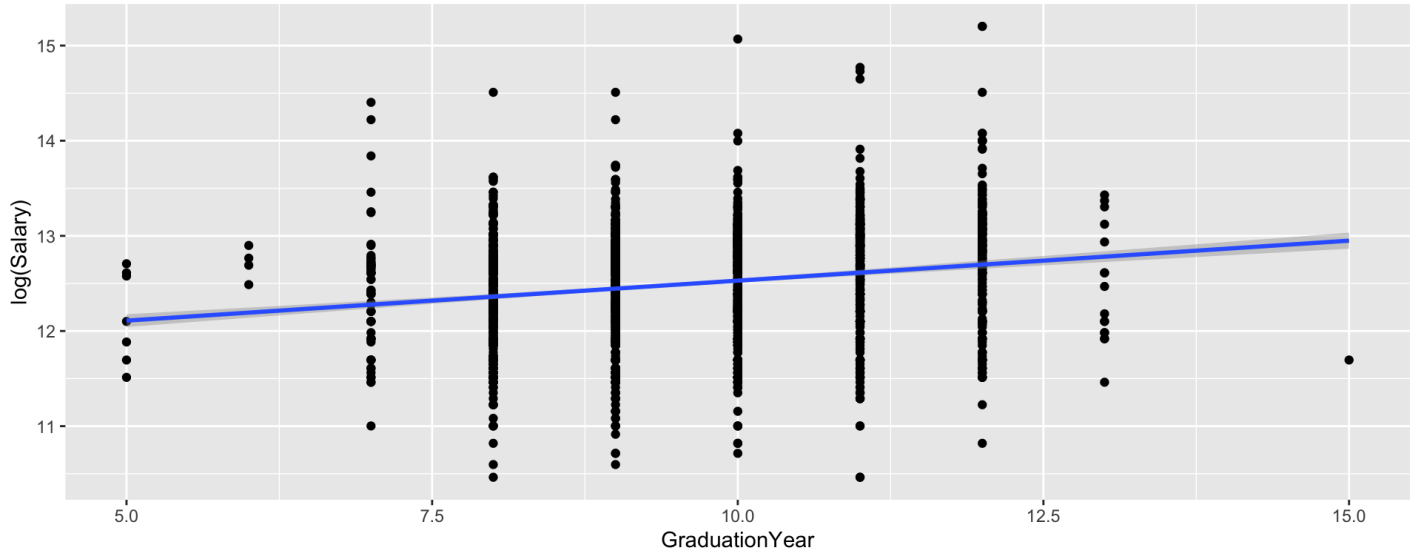
This has a lot of variables that are just basic information, and I decide to remove the useless variables, such as:ID, DOB, 10percentage and so force. For rest of variables, whether or not I use it, it depends on following analysis.

According to graduation year, it is not difficult to calculate how many years the person had graduated. So I do data cleaning in GraduationYear at frist.

```
unique(orignal_data$GraduationYear)
```

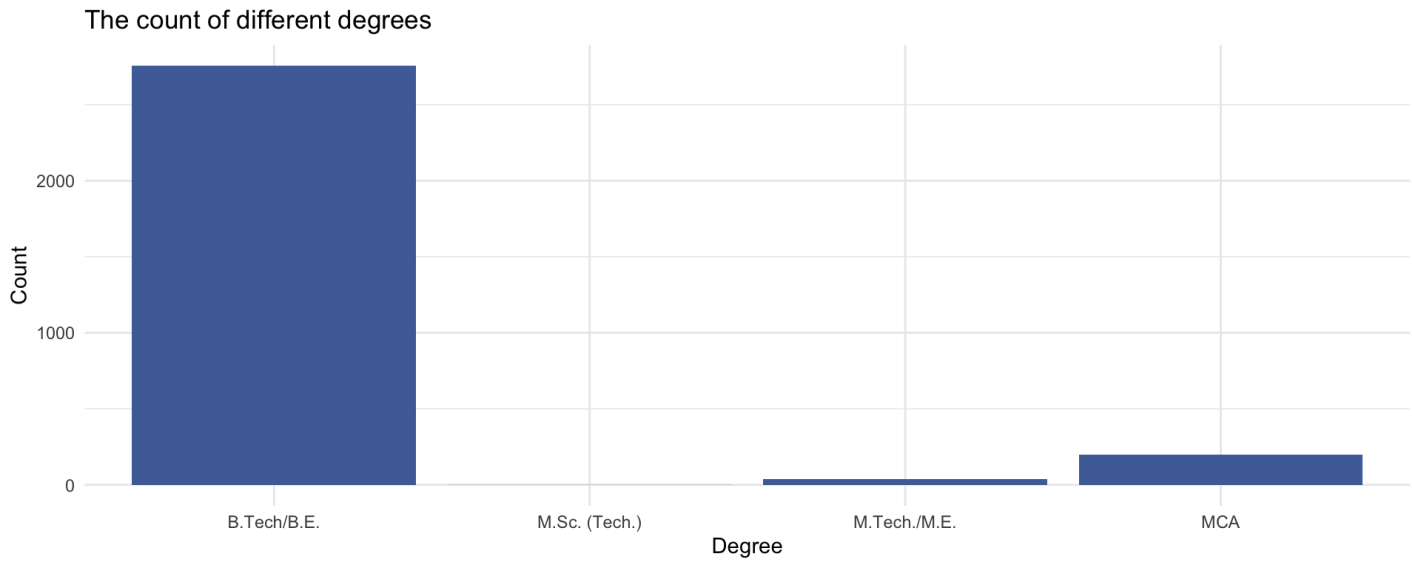
```
##  [1] 2013 2014 2011 2012 2010 2015 2009 2017 2016    0 2007
```

I found an outlier here, 0, so I'm now going to remove it from the data. Secondly, I subtracted the GraduationYear from 2022 to find out how long it took after graduation.

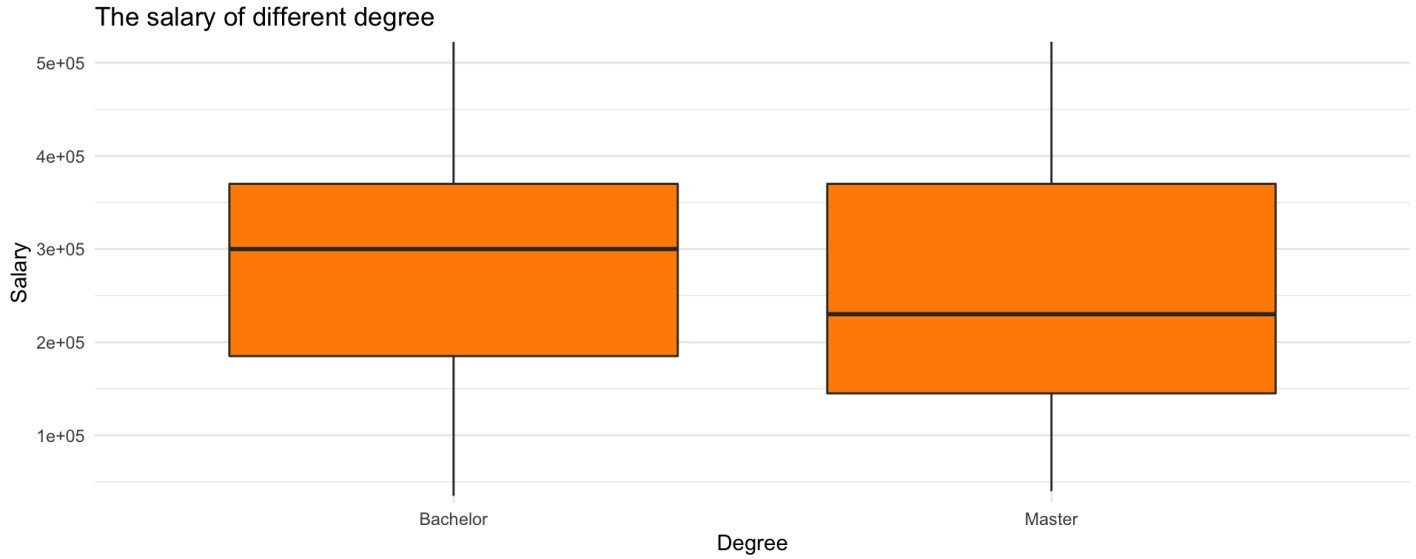


This figure shows the relationship between 'GraduationYear' and 'log(Salary)'. As you can see, when the 'GraduationYear' increases, the 'log(Salary)' adds as well.

As far as I know, the degree is a vital factor to impact what kind of works students can find, it also effect how many salary the students earn. I am going to show how many people there are for different degrees.

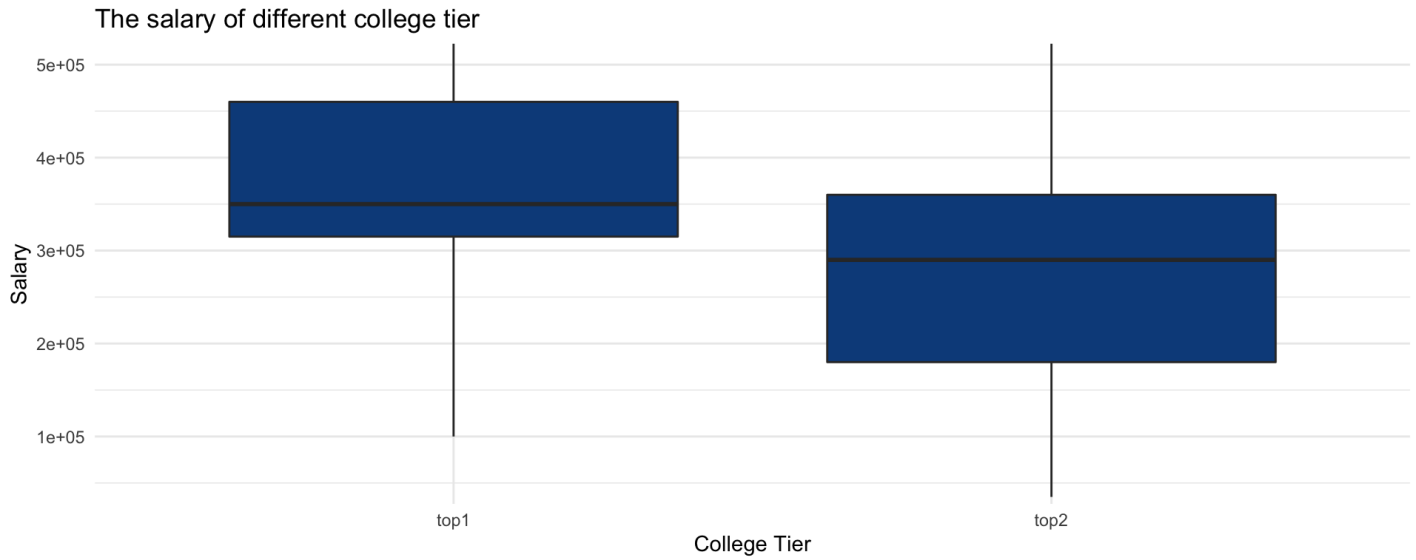


It is easy to see that most people have a bachelor's degree and a small number have a master's degree, so I decided to set the bachelor's degree to 0 and all master's degrees to 1 and then summarize the salary by different degrees.



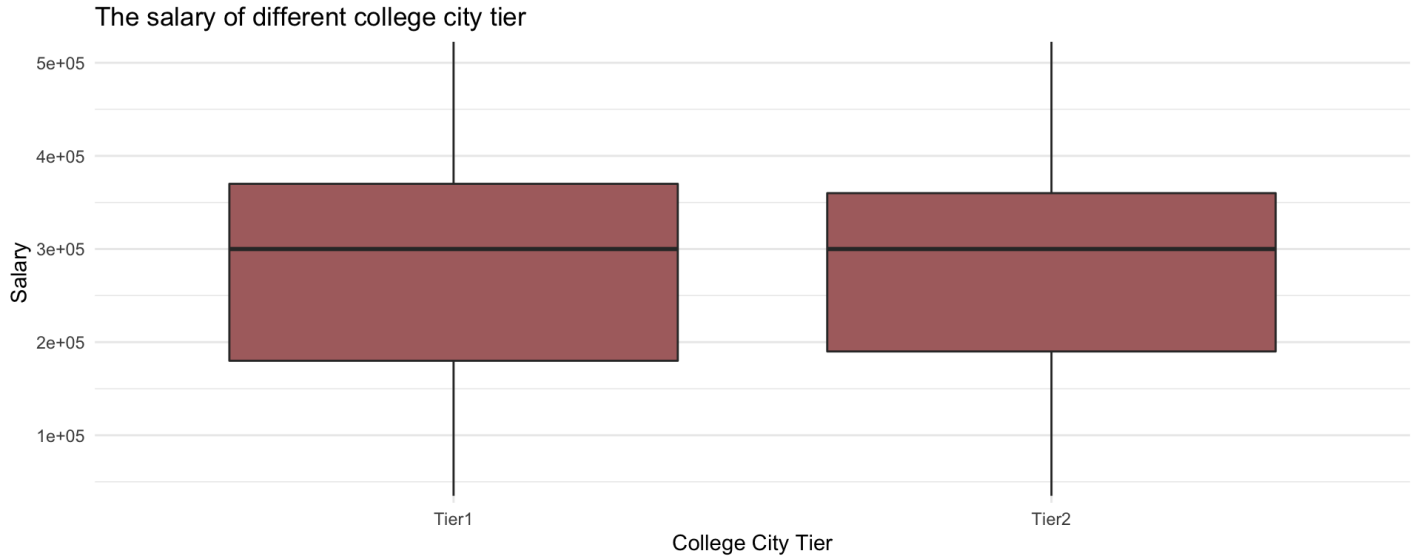
Different degrees do not have a large impact on the overall salary, so I think this degree does not have a large impact on the salary.

Obviously, If a student graduates from a better university, he/she will receive a higher salary after graduation. To prove this point, I used a box line plot with CollegeTier as the variable and salary as the output. The result was plotted as follows.



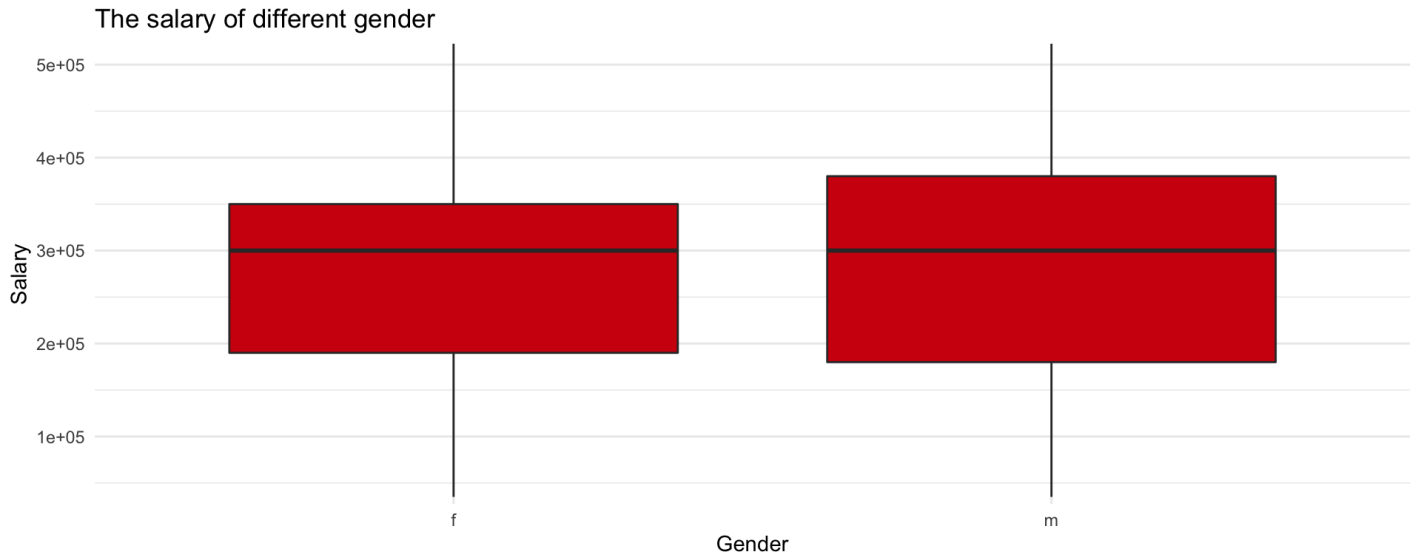
As the figure shows, it is easy to conclude that students graduates from different tier of college have very different salary, the number of '1' means that students graduate from top1 tier college and the number of '2' means that students graduate from top2 tier college. So we can use CollegeTier as a variable to predict salary.

On the other hand, the tier of the city where the student graduated from may also affect the person's situation to find a job with high salary after graduation. In this case, I used the box line plot with CollegeCityTier as the variable and salary as the output as well. The result shows as follows:



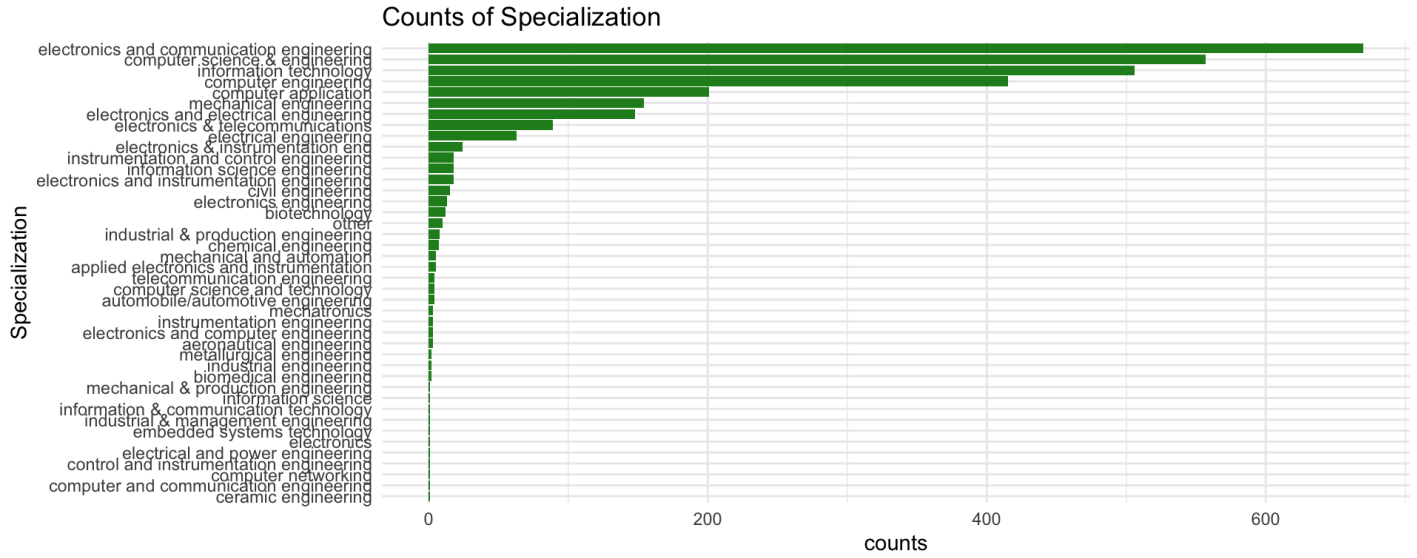
As you can see, we can not find a huge differences between different college city tiers. So I decide to remove this variable.

Besides,I guess that gender is an important variable to determine the person's salary. So I continue using the box line plot to judge whether or not it is a variable to predict the salary. The figure shows as the following:



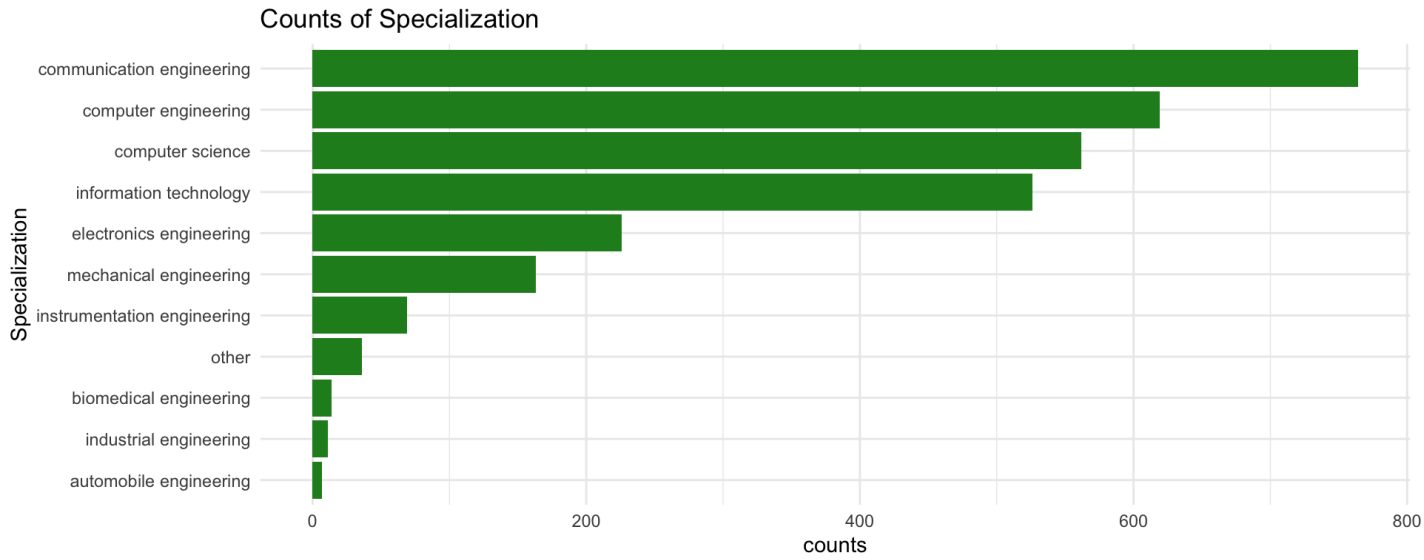
As you can see, there is no obvious difference between two genders. whether male or female, the salary does not change significantly. So I decide to remove this variable.

Specialization is a very important indicator, and different variables have different effects on different specialization. This report focuses how the impact of each variable on salary varies across specializations. I first list all the specializations and calculate their counts.

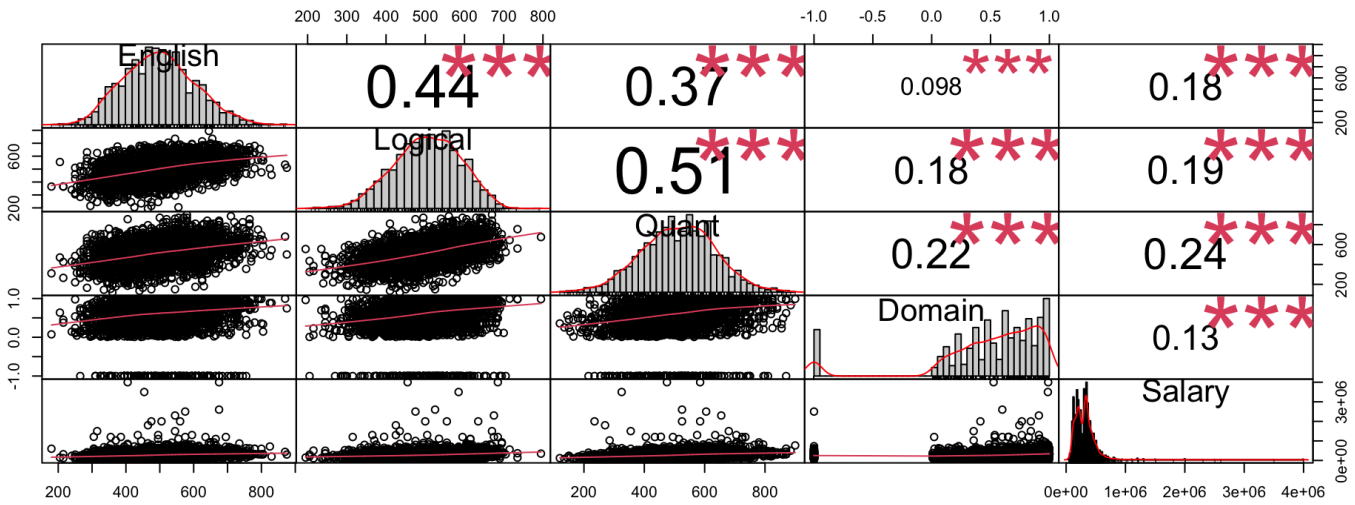


As you can see, there are 42 specializations here, and the small amount of data for some of them is not useful to our analysis. However, I found that many specializations have similar names, such as ‘industrial & management engineering’ and ‘industrial & production engineering’, so I plan to merge them into one specialization.

After data cleaning, I have merged 42 majors into 11 majors, which are ‘computer science’, ‘instrumentation engineering’, ‘information technology’, ‘mechanical engineering’, ‘industrial engineering’, ‘biomedical engineering’, ‘automobile engineering’, ‘electronics engineering’, ‘communication engineering’, ‘computer engineering’, ‘other’. Now, I recalculate their counts, the figure is drawn as follows.



For the next step, I draw the Pearson correlation matrix to select the predictor for the rest of variables, which is score in AMCAT’s different sections.



As the gragh shows, the variables of 'logical' and 'Quant' have relatively high correlation with 'Salary'. So I decide to select both of it as the predictors.

Model fitting

I decide to use multilevel model to fit model. As to selection of variables, I also include 'CollegeTier', 'collegeGPA', 'GraduationYear'. Meanwhile, since 'Salary' is more or less skewed and have heavy tails, I took `log(Salary)` to create new ones.

```
model <- lmer(log(Salary) ~ CollegeTier
              + collegeGPA + GraduationYear
               + Logical + Quant
              + (1 + CollegeTier
              + collegeGPA + GraduationYear
               + Logical + Quant| Specialization)
              ,data = orignal_data_2)
```

```
## boundary (singular) fit: see help('isSingular')
```

Here is the summary of model(fixed effect) and all variables here are considered as statistically significant at $\alpha = 0.5$ level.

	Estimate	Std. Error	t value
(Intercept)	10.7429	0.1974	54.43
CollegeTier	-0.2527	0.0698	-3.62
collegeGPA	0.0077	0.0023	3.23
GraduationYear	0.0781	0.0228	3.42
Logical	0.0004	0.0002	1.64
Quant	0.0012	0.0004	3.18

And the following tables are the summary of random effects, which is random effect of Specialization.

```
##                            (Intercept) CollegeTier collegeGPA GraduationYear
## automobile engineering          0.0759     -0.0009    -0.0005        -0.0154
## biomedical engineering          0.1502     -0.0029    -0.0001        -0.0154
## communication engineering       0.2476     -0.0060     0.0006        -0.0137
## computer engineering           -0.6753      0.0142    -0.0001         0.0552
## computer science                1.4715     -0.0306    -0.0001        -0.1258
## electronics engineering         0.4765     -0.0092    -0.0006        -0.0509
## industrial engineering         -0.6491      0.0118     0.0013         0.0787
## information technology         -0.0130     -0.0015     0.0012         0.0239
## instrumentation engineering    -0.1453      0.0039    -0.0006         0.0000
## mechanical engineering         -0.8057      0.0164     0.0003         0.0782
## other                          -0.1334      0.0048    -0.0014        -0.0150
##                            Logical  Quant
## automobile engineering       1e-04  1e-04
## biomedical engineering       0e+00  0e+00
## communication engineering   -1e-04 -2e-04
## computer engineering         2e-04  1e-04
## computer science            -4e-04 -2e-04
## electronics engineering      0e+00  0e+00
## industrial engineering       0e+00 -2e-04
## information technology      -2e-04 -3e-04
## instrumentation engineering  1e-04  2e-04
## mechanical engineering       1e-04  1e-04
## other                        2e-04  4e-04
```

Result

Interpretation

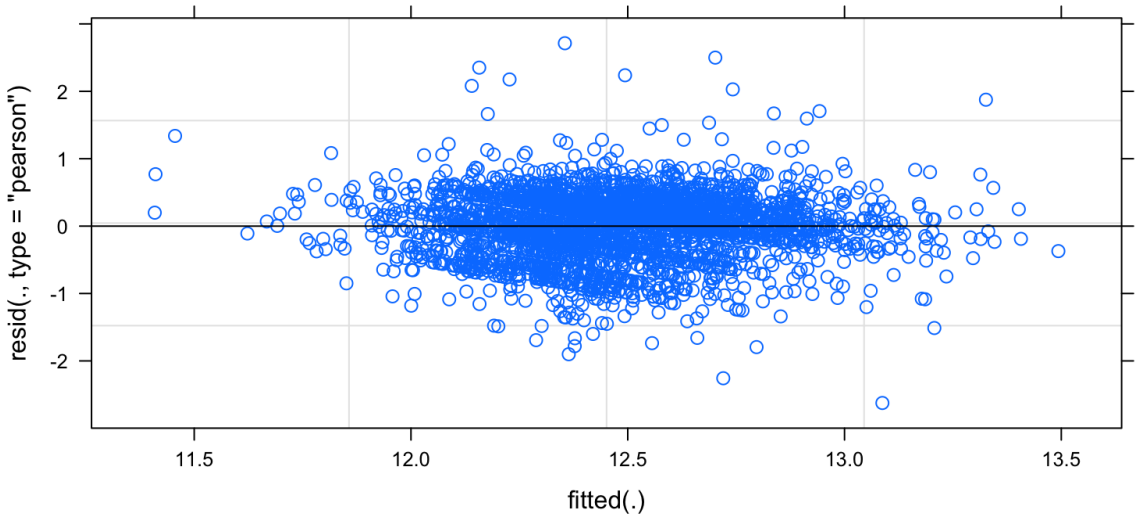
Let's take 'computer science' for an example. We are able to get the following formula of fixed effect:

$$log(Salary) = 1.4715 - 0.0206 \times CollegeTier - 0.0001 \times collegeGPA - 0.1258 \times GraduationYear - 0.0004 \times Logical - 0.0002 \times Quan$$

As the table shows, different specialization have huge differences, some part of coefficients are positive but another part are negative. In this case, we need to focus on a special specialization to research it variables.

Discussion

Model checking



The plot is residual plot. According to it, the mean value of residuals is approximately 0. Yet as the fitted value close to 0, there's no negative residuals.

Reference

[1] Micheal, C. *Mixed Models with R*. https://m-clark.github.io/mixed-models-with-R/ (https://m-clark.github.io/mixed-models-with-R/)

[2] Ayman Siraj. RPubs. Weblog. https://rpubs.com/aymansir/usflightdelay (https://rpubs.com/aymansir/usflightdelay)

[3] Micheal, P. *Chapter 18: Testing the Assumptions of Multilevel Models*. https://ademos.people.uic.edu/Chapter18.html (https://ademos.people.uic.edu/Chapter18.html)

Appendix

Variable distributions