In Class EDA Activity

Jing Xu

2022-10-04

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Read data

The data is how much graduate students with engineering degree earn. We will talk about the origin of the data after your submission.

```r
RawData<- read.csv("Engineering_graduate_salary_Simple.csv")
```

Here are the description of the variables included:

- 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 which the candidate attended for her/his undergraduate
- CollegeTier: Each college has been annotated as 1 or 2. The annotations have been computed from the average AMCAT scores obtained by the students in the college/university. Colleges with an average score above a threshold are tagged as 1 and others as 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. This is annotated based on - the population of the cities.
- CollegeState: Name of the state in which the college is located
- GraduationYear: Year of graduation (Bachelor's degree)

Clean Data

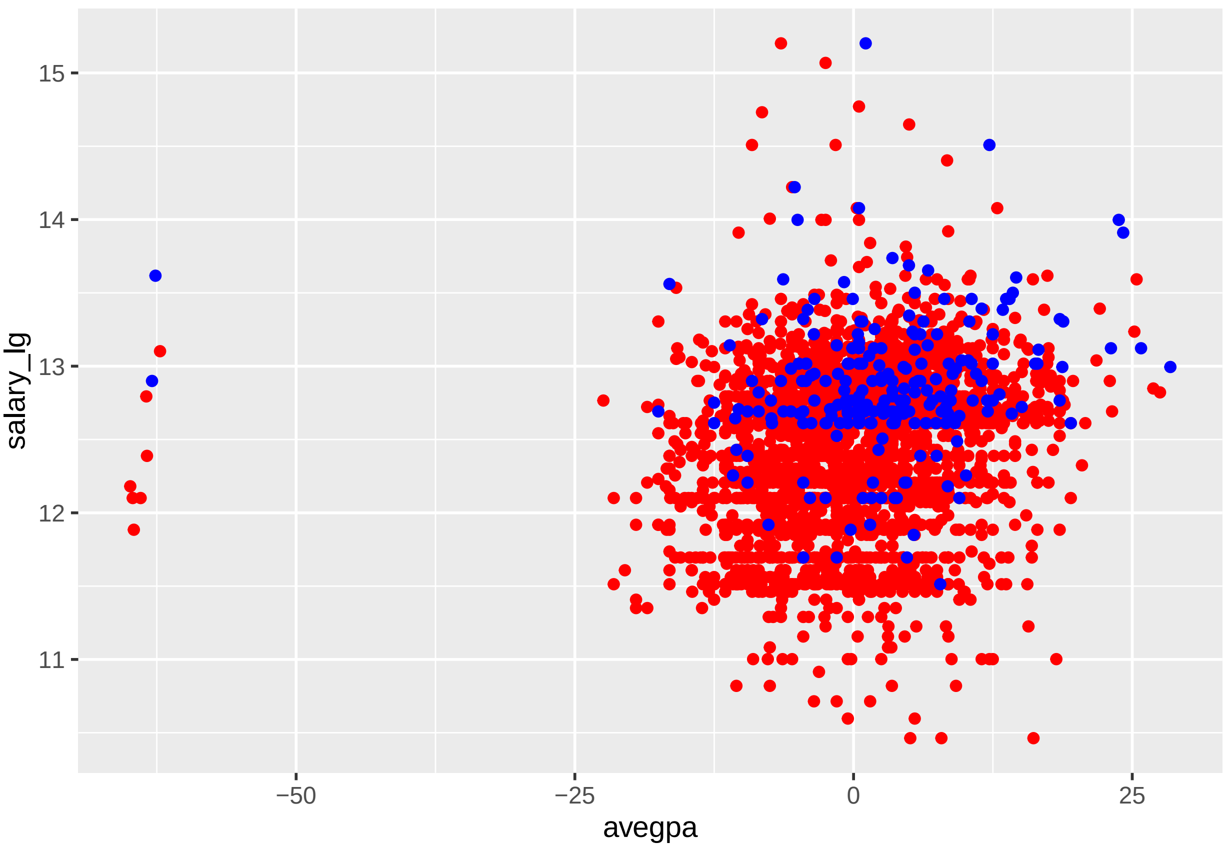
```
RawData$male <- 0
RawData$g_college <- 0
RawData$master <- 0
RawData$avegpa <- 0
RawData$salary_lg <- 0
g <- mean(RawData$collegeGPA)
for (i in RawData$X){
  if (RawData$Gender[i] =="m")
    RawData$male[i] <- 1
  if (RawData$CollegeTier[i] ==1)
    RawData$g_college[i] <- 1
  if (RawData$Degree[i] == "MCA")
    RawData$master[i] <- 1
  RawData$avegpa[i] <- RawData$collegeGPA[i] - g
  RawData$salary_lg[i] <- log(RawData$Salary[i])
}
```

EDA

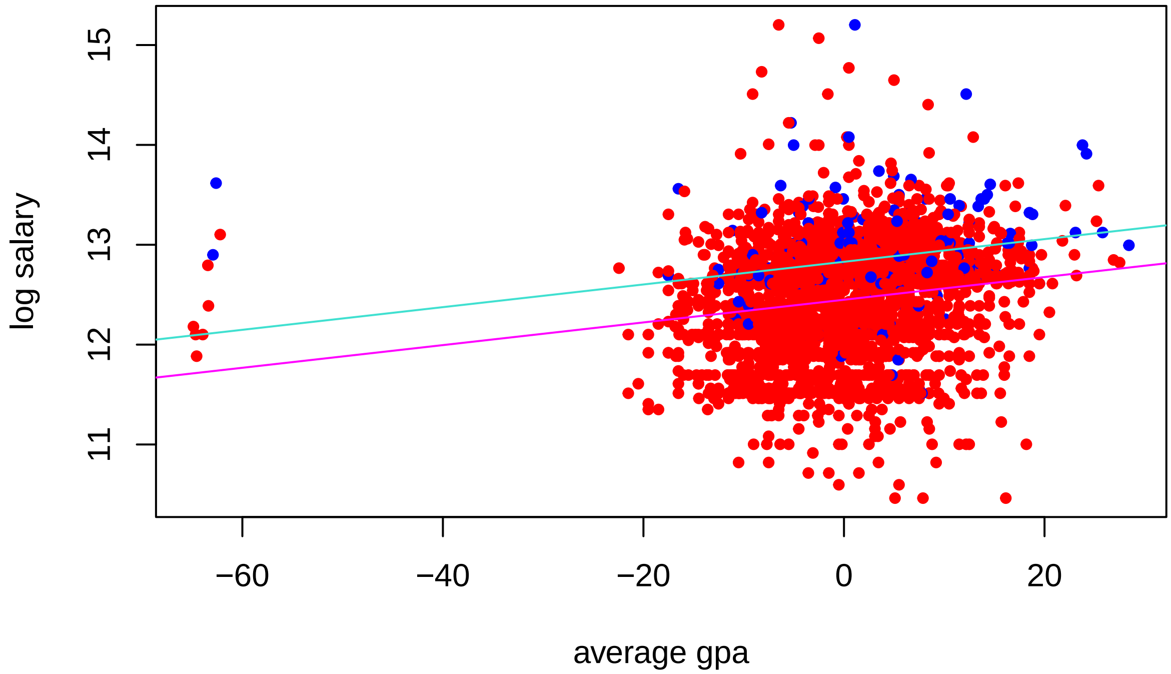
```
fit_1 <- lm(salary_lg ~ g_college + avegpa, data = RawData)
summary(fit_1)
```

```
##
## Call:
## lm(formula = salary_lg ~ g_college + avegpa, data = RawData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16944 -0.31400  0.07135  0.33329  2.82639
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 12.449316   0.010009 1243.856   <2e-16 ***
## g_college    0.380028   0.036586   10.387   <2e-16 ***
## avegpa       0.011352   0.001189    9.544   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5268 on 2995 degrees of freedom
## Multiple R-squared:  0.06793,    Adjusted R-squared:  0.06731
## F-statistic: 109.1 on 2 and 2995 DF,  p-value: < 2.2e-16
```

```
c <- coef(fit_1)
ggplot(RawData,aes(avegpa,salary_lg))+
  geom_point(data=RawData %>% filter(g_college==0), col = "red")+
  geom_point(data=RawData %>% filter(g_college==1), col = "blue"
             )
```



```
color <- ifelse(RawData$g_college==0,"red", "blue")
plot(RawData$avegpa,RawData$salary_lg,pch=20, col = color, xlab = 'average gpa', ylab = 'log salary')
abline(c[1], c[3],col="magenta")
abline(c[1] + c[2], c[3],col="turquoise")
```



```
#people who get their master degree is averagely earning .38% higher salary comparing to those who didn
#1 point higher aggregate GPA at one's graduation is averagely associated with a increase of .011 perce
```