Biostatistics Project

Mia (Wei-Jhen Suen)

2024.06.02

I. Introduction

Inspired by a recent podcast about the gender-salary relationship, I would like to delve deeper into the relationship between salary and different factors. In this project, my focus is to uncover and interpret patterns in the data in order to understand the reasons influencing salary levels. I will be analyzing a salary-based dataset from Kaggle, originating from an anonymous tech company. (Unfortunately, I can't attach the link here because the original source website seems to be no longer available…)

II. Look at the data

To begin with, I read in the csv. file from my project's directory and check out the data structure.

```
#> spc_tbl_ [6,704 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
#>  $ Age                : num [1:6704] 32 28 45 36 52 29 42 31 26 38 ...
#>  $ Gender             : chr [1:6704] "Male" "Female" "Male" "Female" ...
#>  $ Education Level    : chr [1:6704] "Bachelor's" "Master's" "PhD" "Bachelor's" ...
#>  $ Job Title          : chr [1:6704] "Software Engineer" "Data Analyst" "Senior Manager" "Sales Asso
#>  $ Years of Experience: num [1:6704] 5 3 15 7 20 2 12 4 1 10 ...
#>  $ Salary             : num [1:6704] 90000 65000 150000 60000 200000 55000 120000 80000 45000 110000
#>  - attr(*, "spec")=
#>   .. cols(
#>   ..   Age = col_double(),
#>   ..   Gender = col_character(),
#>   ..   `Education Level` = col_character(),
#>   ..   `Job Title` = col_character(),
#>   ..   `Years of Experience` = col_double(),
#>   ..   Salary = col_double()
#>   .. )
#>  - attr(*, "problems")=<externalptr>
```

There are six columns - *Age*, *Gender*, *Education Level*, *Job Title*, *Years of Experience*, and *Salary* along with 6.704 data in total. I noticed that the data types of *Gender*, *Education Level*, and *Job Title* are characters. It reminds me of the previous course materials that it's crucial to convert a vector object to a factor, especially while analyzing data. It helps with a fixed set of acceptable values, avoiding errors from any arithmetic operations on that column. Therefore, I overwrite *Gender*, *Education Level*, and *Job Title* as factors.

Then, let's take a look at the summary of the data.

```
#>       Age           Gender              Education Level
#>  Min.   :21.00   Female:3014   Bachelor's Degree:2267
#>  1st Qu.:28.00   Male  :3674   Master's Degree  :1573
#>  Median :32.00   Other :  14   PhD              :1368
#>  Mean   :33.62   NA's  :   2   Bachelor's       : 756
#>  3rd Qu.:38.00                 High School      : 448
#>  Max.   :62.00                 (Other)          : 289
#>  NA's   :2                     NA's             :   3
#>                       Job Title    Years of Experience     Salary
#>  Software Engineer        : 518   Min.   : 0.000      Min.   :   350
#>  Data Scientist           : 453   1st Qu.: 3.000      1st Qu.: 70000
#>  Software Engineer Manager: 376   Median : 7.000      Median :115000
#>  Data Analyst             : 363   Mean   : 8.095      Mean   :115327
#>  Senior Project Engineer  : 318   3rd Qu.:12.000      3rd Qu.:160000
#>  (Other)                  :4674   Max.   :34.000      Max.   :250000
#>  NA's                     :   2   NA's   :3           NA's   :5
```

For *Gender*, the number of *Male* is slightly more than *Female* but not a huge gap. For *Education Level*, I notice some overlap naming can be combined together, which I will process afterward. As for *Salary*, the minimum is pretty low, while the maximum is actually not as high as I imagine. Driven by curiosity, I checked the job title of the highest salary - it's from the company's CEO.

Problem shooting in this section

Problem 1: It is not allowed to have space in columns' names.
Solution: I tried many traditional ways, like substitute the space with dot(.) or underline(_), but the results is not as satisfied as I wished. At the end, I found an elegant but useful way by simply adding ` ` to the names in the code, so the errors can be solved without changing original data.

Problem 2: After removing NA rows at the beginning, there are still NA while doing visualization.
Solution: The professor pointed out that it's not really a good way to remove all the rows with NA at the very beginning, cause it also remove other data in those rows which may be useful while analyzing specific columns. Therefore, I decided to wrangle the data individually for different purposes, and remove NA at the very end to reserve as much data as possible.

III. Analyze the data

After familiarizing myself with the fundamental structure of the data, my goal is to delve deeper into the relationship between salary and other factors from three aspects:

A. Gender v.s. Salary

B. Education Level v.s. Salary

C. Years of Experience v.s. Salary (based on Education Level)

In the following sections, I will:

1. Wrangle the data
2. Compute summary statistics
3. Visualize the data

4. Check assumptions
5. Interpret the results

A. Gender v.s. Salary

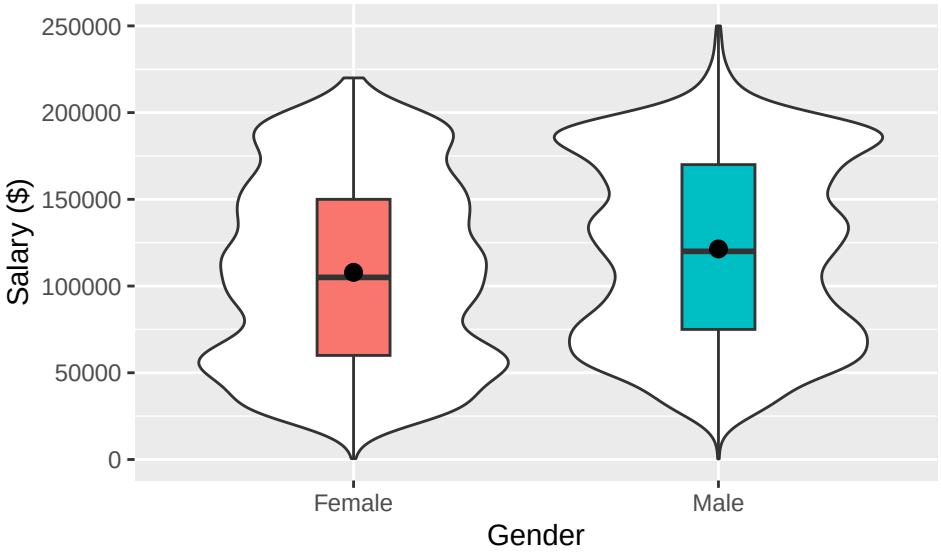
In this section, I want to find out whether gender has an impact on salaries within the data and to uncover any noteworthy patterns or differences in earnings based on gender.

First, I wrangle the data by selecting the columns *Gender* and *Salary*, and filtering only the *Female* and *Male* rows which I want to look into. Last but not least, omit NA rows, and.. data is ready!

```
#> # A tibble: 2 x 4
#>   Gender     N    mean     sd
#>   <fct>  <int>   <dbl>  <dbl>
#> 1 Female  3013 107889. 52724.
#> 2 Male    3672 121390. 52093.
```

Take a look at the summary statistics - counts, means and standard deviations. For the average income, men earn $13.500 more than women per year. Also, men have a bigger variance for salary's distribution, while women have a more centered one.

For data with two or more numerical variables, I use the violin-boxplots to depict both summary statistics (from boxplots) and the density of two categories (from violin plots).



According to the plot, we can conclude that men have a higher density at higher payments $185.000 approximately. Also, men have a higher payment peak, earning up to $250.000, while women can only earn a maximum $220.000 per year.

```
#>
#>  F test to compare two variances
#>
#> data:  Salary by Gender
#> F = 1.0244, num df = 3012, denom df = 3671, p-value = 0.4878
#> alternative hypothesis: true ratio of variances is not equal to 1
```

```
#> 95 percent confidence interval:
#>  0.9569885 1.0967467
#> sample estimates:
#> ratio of variances
#>           1.024368


#>
#>  Two Sample t-test
#>
#> data:  Salary by Gender
#> t = -10.486, df = 6683, p-value < 2.2e-16
#> alternative hypothesis: true difference in means between group Female and group Male is not equal to
#> 95 percent confidence interval:
#>  -16024.79 -10976.95
#> sample estimates:
#> mean in group Female   mean in group Male
#>             107889.0             121389.9
```

The p-value of F-test is p = 0.4878, greater than the significance level 0.05, which means there is no significant difference between the variances of the two genders. Therefore, we can use the classic t-test which assume equality of the two variances.

For two statistically independent samples, I use Two sample t-test to determine if salary means for two genders are equal. As you can see, t statistic value is t = -10.486, and degrees of freedom is df = 6683. The p-value of the t-test is super small, less than the significance level 0.05. We can conclude that men's average salary is significantly different from women's. The confidence interval of the mean at 95% is [-16024.79, -10976.95].

```
#> Cohen's d |         95% CI
#> -------------------------
#> -0.26     | [-0.31, -0.21]
#>
#> - Estimated using pooled SD.


#> t(6683) = -10.49, p < .001, d = -0.26 [-0.30; -0.21]
```

Based on Cohen's d, the effect size -0.26 is rather small, indicating that the salary difference between male and female is not practically significant.

Overall, there's no significant impact of gender on salaries within this tech company dataset.

B. Education Level v.s. Salary

Since the steps for the following sections are pretty similar to the first one, I will skip some similar details and focus only on the main description and interpretation.

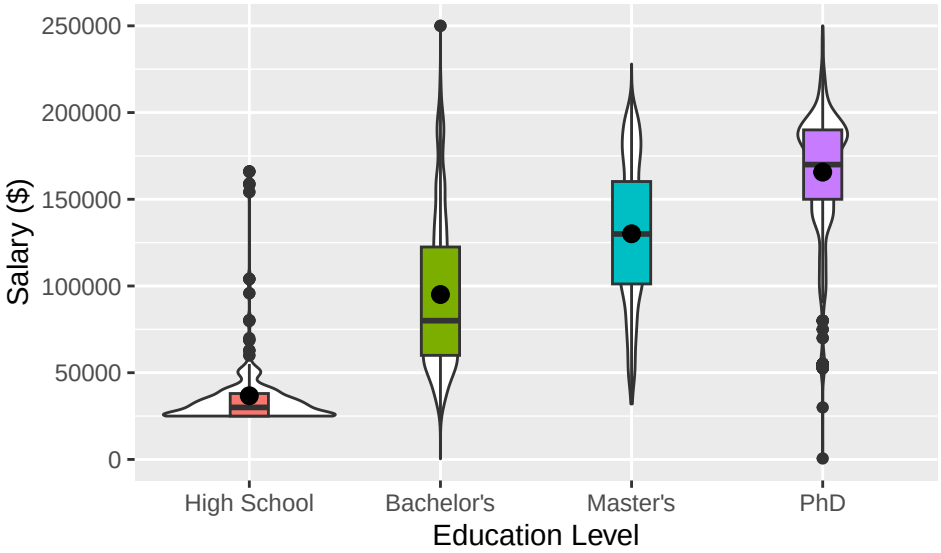
In this section, I want to find out whether education levels have impacts on salaries within the data and to uncover any noteworthy patterns or differences in earnings based on education levels.

```
#> [1] "Bachelor's"        "Bachelor's Degree" "High School"
#> [4] "Master's"          "Master's Degree"   "phD"
#> [7] "PhD"
```

In the column *Education Level*, there are different naming for same education levels, so I rename the factors using `mutate` function.

```
#> # A tibble: 4 x 4
#>   `Education Level`     N    mean     sd
#>   <fct>             <int>   <dbl>  <dbl>
#> 1 High School         448  36707. 22549.
#> 2 Bachelor's         3021  95083. 44092.
#> 3 Master's           1860 130112. 40641.
#> 4 PhD                1369 165651. 34340.
```

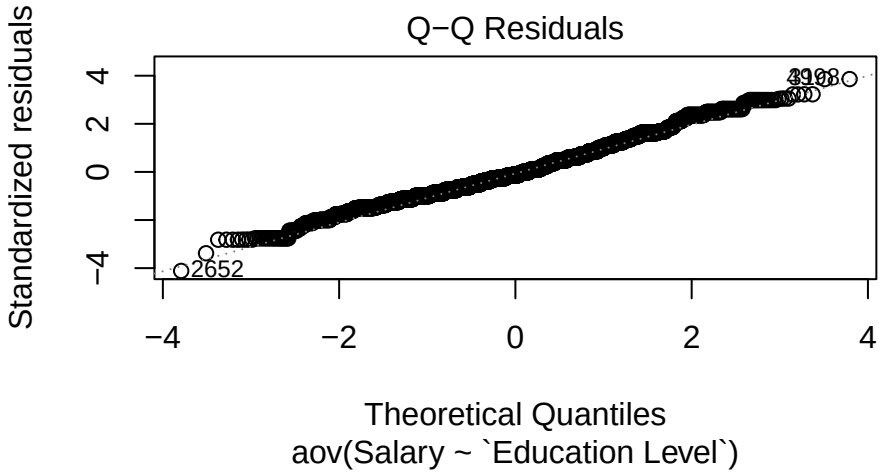
We can tell from the means that the average salary will increase with education level.




```
#> Analysis of Variance Table
#>
#> Response: Salary
#>                     Df     Sum Sq    Mean Sq F value    Pr(>F)
#> `Education Level`    3 7.8809e+12 2.6270e+12  1630.9 < 2.2e-16 ***
#> Residuals         6694 1.0782e+13 1.6107e+09
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#>                     Df    Sum Sq   Mean Sq F value Pr(>F)
#> `Education Level`    3 7.881e+12 2.627e+12    1631 <2e-16 ***
#> Residuals         6694 1.078e+13 1.611e+09
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I run the between-subjects ANOVA using two different functions - `lm()` and `aov()`, and turn out they have exactly same results. As the p-value is much less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with "*" in the summary.



As all the points fall approximately along this reference line, we can assume normality.

```
#> $emmeans
#>  Education Level emmean   SE   df lower.CL upper.CL
#>  High School      36707 1896 6694    32990    40424
#>  Bachelor's       95083  730 6694    93652    96514
#>  Master's        130112  931 6694   128288   131936
#>  PhD             165651 1085 6694   163525   167778
#>
#> Confidence level used: 0.95
#>
#> $contrasts
#>  contrast                 estimate   SE   df t.ratio p.value
#>  High School - Bachelor's   -58376 2032 6694 -28.730  <.0001
#>  High School - Master's     -93405 2112 6694 -44.222  <.0001
#>  High School - PhD         -128945 2184 6694 -59.028  <.0001
#>  Bachelor's - Master's      -35029 1183 6694 -29.614  <.0001
#>  Bachelor's - PhD           -70569 1308 6694 -53.969  <.0001
#>  Master's - PhD             -35539 1429 6694 -24.867  <.0001
#>
#> P value adjustment: bonferroni method for 6 tests
```

Post-hoc test is done to identify which groups differ from each other.

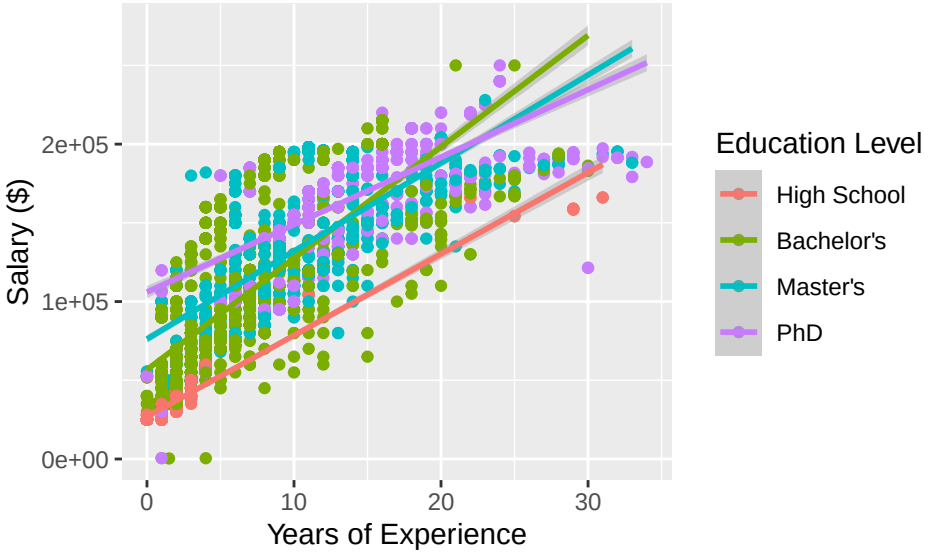
```
#> # Effect Size for ANOVA (Type I)
#>
#> Parameter       |    ² |       95% CI
#> --------------------------------------
#> Education Level | 0.42 | [0.41, 1.00]
#>
#> - One-sided CIs: upper bound fixed at [1.00].
```

The effect size $^2 = 0.42$ represents a moderate effect, indicating that the difference between groups has no practical significance, but still has a reference value.

C. Years of Experience v.s. Salary (based on Education Level)

In this section, I want to observe the salary trend based on different education level's working experience.

```
#> `geom_smooth()` using formula = 'y ~ x'
```



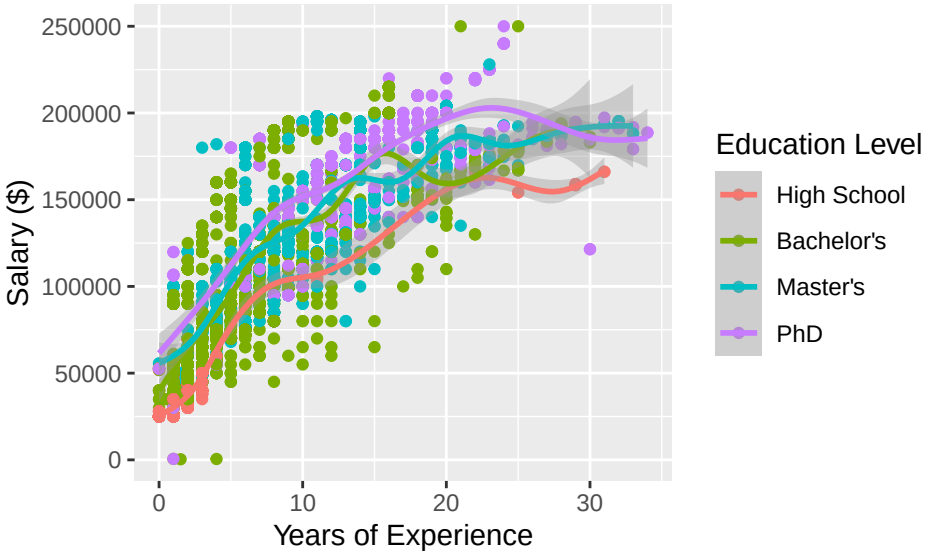
From my perspective, this is a very interesting graph.

Starting salaries increase with the level of education, which aligns with many people's expectations. Interestingly, as years of work experience grow, the salary for bachelor's degrees experiences the highest significant rise, followed by master's degrees, and PhD the least. Salaries for those with high school education remain consistently the lowest.

However, the previous analysis indeed showed that salaries increase with higher levels of education. Why does it seem contradictory to this result now? I infer maybe it's because most employees in this tech companies have lower seniority, not yet surpassing the intersection point on the salary graph.

2024.04.09 Update:
After reading the professor's feedback, I think he made a really good point. From the linear model, it's hard to tell if the Bachelor's line is influenced by a few specific cases. Maybe the CEO of the company has a Bachelor's degree, which influences the trend of the line. He recommended me to revise the visualization a bit in order to fit "flexible curves" to the data. From the following graph, We can tell that the Bachelor's line is still below Master's in most cases. In conclusion, people with higher degrees have higher salaries "on average".

```
#> `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
#>
#> Call:
#> lm(formula = Salary ~ `Years of Experience`, data = dat3_1)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -141962  -22570   -4207   20881   95793
#>
#> Coefficients:
#>                       Estimate Std. Error t value Pr(>|t|)
#> (Intercept)           64295.18     673.17   95.51   <2e-16 ***
#> `Years of Experience`  6637.23      64.73  102.53   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 30380 on 6248 degrees of freedom
#> Multiple R-squared:  0.6272, Adjusted R-squared:  0.6272
#> F-statistic: 1.051e+04 on 1 and 6248 DF,  p-value: < 2.2e-16
```

For the linear model, I exclude the education factor and focus only on the regression of salary and work experience.

For linearity and homogeneity, the plot suggests they are not perfect at the tail end. Perhaps it's because there are fewer data, but except for that, they look pretty good.

For normality of residuals, the plot does suggest that the residuals might not be normal, so I check this with `check_normality()` which runs a Shapiro-Wilk test.

```
#> Warning: Non-normality of residuals detected (p < .001).
```

The result confirms that the residuals are not normally distributed.

```
#>
#>      Multiple regression power calculation
#>
```

```
#>               u = 1
#>               v = 35
#>              f2 = 0.224397
#>       sig.level = 0.05
#>           power = 0.8
```

Challenges in this section

Challenge 1: It took me quite some of time to determine the most suitable plots for different data types and select appropriate tests to assess assumptions.

Challenge 2: I haven't really figure out how to calculate the parameters for effect size, and also the format in apa package. The interpret for numerical statistics requires solid statistical knowledge, with which I haven't been acquainted enough. In my opinion, these are the most crucial steps for data analysis, and it can only be overcame and enhanced by consistent practice and familiarity, reaching out to multiple information and allowing for greater proficiency over time.

IV. Conclusion

In this report, I analyze a dataset of 6,704 employees in a tech company, comparing their salaries with various factors. The results indicate that gender does not have a significant impact on salaries, while education level and years of work experience are positively correlated with average wages.

I'm really happy to learn R language and analytical skills from the beginning in this semester, and apply them to this report. Although the process was quite challenging, solving each problem brought a huge sense of accomplishment to me. Thankfully, the results are satisfying and intriguing as well. I plan to continue exploring how to interpret data and will try to present the results in APA format to enhance and fully complete this report.

V. Reference

1. https://spressi.github.io/biostats/
2. https://psyteachr.github.io/quant-fun-v2/index.html
3. http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r
4. http://www.sthda.com/english/wiki/one-way-anova-test-in-r
5. http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/