STAT340 Final Report: Gender Base Salary Gap in STEM Field

12/02/2021

Group Member:

- Cecheng Chen(cchen549)
- Zhuocheng Sun(zsun273)
- Boya Zeng (bzeng7)
- Yueyu Wang(wang2537)
- Zihan Zhu (zzhu338)

Abstract

Today, women earn approximately 82 cents for every dollar earned by a man(Carlton, G., 2021, The biggest barriers for women in STEM). However, a report from Scientific American shows that in the STEM field, males and females have approximately equal average base salaries (Ceci et al., 2015, Scientific American). Unlike other working fields, it seems that the gender salary gap in STEM fields is the least obvious. However, during our research and feedback from our friends, it seems that there still exists some salary gap in gender. Also, we found that in the STEM field, other factors such as regions, employee' education backgrounds, etc. also seem correlated with base salaries. Therefore, we want to figure out whether there truly exists gender differences in the STEM field, and if there exists, does such difference correlate with other factors like regions, job positions, different kinds of companies, etc? Based on these questions, our data is mainly focusing on the different personal situations and different salaries of employee in the STEM fields. Based on our research, we find that there still exists gender base salary gap in STEM field, and such gender base salary gap also depends on factors like regions (in state level), Job titles (positions) and Companies.

Dataset descriptions

We totally use two data sets for our program

Dataset1: Data Science and STEM Salaries

The URL for this data set: https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries

This data set was scraped off levels.fyi by Jack Ogozaly. levels.fyi is a website that lets you compare career levels & compensation packages across different tech companies, and is generally considered more accurate in terms of actual tech salaries relative to other compensation sites like glassdoor.com. We use this dataset because this dataset has most information we want: The base salary, gender, company, location, races, etc of an employee. We can easily use these information to compare median base salary differences in gender and compare such difference based on other factors. With this data set, we can answer our research questions more easily and efficiently.

Description of this dataset: This dataset contains 62,000 salary records from top STEM companies like Amazon, Apple, Google, SpaceX, etc. For each salary record, it contains the company employee works for, job titles and positions, base salary, total year salary, gender, race, and other useful information. We can use this data set to analyze the income situation of workers in these companies based on the characteristics of personal and companyies.

Variables :

- Timestamp: When the data was recorded.
- Company: The company name where the employee works (Google, Facebook, etc)
- Level: What level the employee is at.
- Title: Role title.
- Total yearly compensation: Total yearly compensation.
- Location: Job location.
- Years of experience: Years of Experience.
- Years at company: Years of experience at said company.
- Tag: Job type
- Base salary: Base salary an employee earned in a year
- Stock grant value: The equivalent value for the stocks the employee received.
- Bonus: These bonuses could be in the form of a lump sum cash payment, increment cash payments, stock options, or even an added vacation, we only count the bonus in dollar amounts here.
- Gender: gender identity of employee (male, female or other)
- Other details: Other details for the employee
- City id: The id for the city where the employee works
- Dmaid: Designated Market Areas (DMAs) delineate the geographic boundaries of 210 distinctive regions to assess TV penetration of audience counts within the U.S. for a viewership year.
- Row Number: row number of the data entry
- Masters_Degree: Whether the employee has a Master Degree (1: Yes, 0: No)
- Bachelors_Degree: Whether the employee has a Bachelor Degree (1: Yes, 0: No)
- Doctorate_Degree: Whether the employee has a Doctor Degree (1: Yes, 0: No)
- Highschool: Whether the employee has a High school Degree (1: Yes, 0: No)
- Some_College: if the employee has education limited to some college education.
- Race_Asian: if the employee is asian (1: Yes, 0: No)
- Race_White: if the employee is white (1: Yes, 0: No)
- Race_Two_Or_More: if the employee has two or more races. (1: Yes, 0: No)
- Race_Black: if the employee is black (1: Yes, 0: No)
- Race_Hispanic: if the employee is hispanic (1: Yes, 0: No)
- Race: Racial identity of the employee
- Education: the Education level of employee

Dataset2: Cost of Living Index by State 2021

The URL for this data set: https://worldpopulationreview.com/state-rankings/cost-of-living-index-by-state

This data set is from and gathered by World Population Review website that allows you to compare the cost of living index for different states. We used this dataset because for some states like CA, WA, NY and MA are very expensive states and many tech companies are based there, so base salaries will be higher there, and that may nagetively impact our regression model if we use state as one possible predictor. Therefore, to best predict base salary of different states and minimize such effects, we plan to add the cost of living index by state variable as an addition predictor.

Description of this dataset: This data set contains the cost-of-living-index for each state in America, also contains the cost index in sub living categories like Grocery, Housing, Utilities, etc.

Variables:

- Cost Index: The overall cost of living index for each state in America, the higher the index is, the higher overall living expense in that state.
- Grocery: The cost of index in Grocery category for each state in America
- Housing: The cost of index in Housing category for each state in America
- Utilities: The cost of index in Utilities category for each state in America
- Transportation: The cost of index in Transportation category for each state in America
- Misc: The cost of index in misc category for each state in America

Statistical Questions:

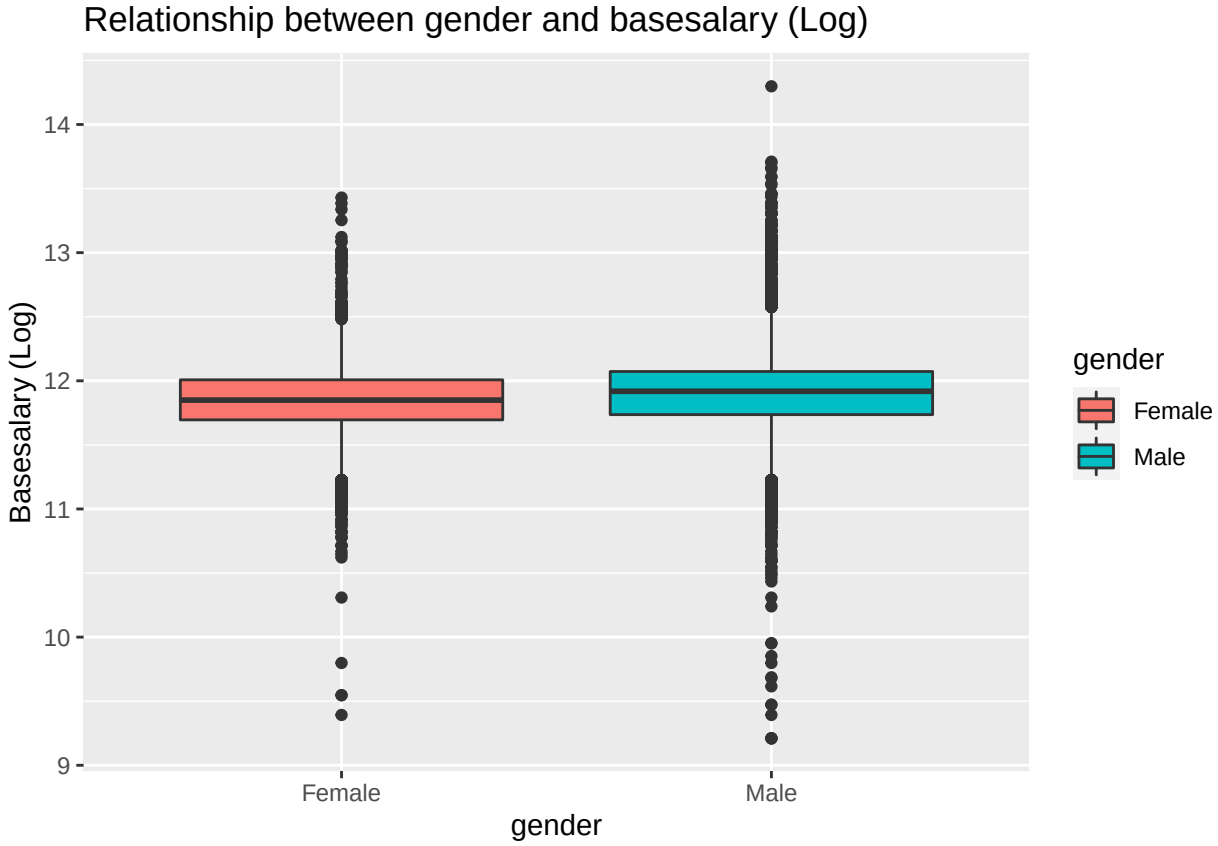
Main question: Is there a gender gap in income level in different data science and STEM subfields and different regions? Does the difference vary depending on other factors (e.g., education, subfields, companies, regions, etc.)?

Process and summaries:

Because in our data set may contain outliners (i.e Some CEOs and employees may have extremely high base salaries), so instead of using average base salary, our research focus on the median base salaries.

At first, we want to consider whether there truly exists gender gap in base salary, so we drow a box plot to show the median income (in logarithm) of males and females.

```
## Warning: Removed 823 rows containing non-finite values (stat_boxplot).
```



According to the above plot, It seems that the median income of females is only slightly lower than males income. Therefore, we cannot tell whether males and females truly have different median income in stem fields, so we conduct hypothesis tests to verify what is shown in the plot.

Our null hypothesis is $H_0 : Median\_Salary_{male} = Median\_Salary_{female}$

The alternative hypothesis is $H_1 : Median\_Salary_{male} \neq Median\_Salary_{female}$

First, we conduct a two sample wilcoxon test, a test focus on testing the median of datasets, and the test result shows as follows:

```
##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  basesalary by gender
## W = 78023286, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Based on the p value of the test result, which is smaller than 2.2e-16, which is much smaller than 0.01 and is highly significant. We have strong statistical evidence to reject the null hypothesis in favor of the conclusion that the median income for males and females are not the same.

We also apply the Monte carol testing to test the null hypothesis. Firstly, we randomly assign the income data into male group and female groups, and compare their medians, store the median income difference as an element in a list Replicate . Then we repeat the step mentioned above 1000 times and use all elements in Replicate to generate a 95% confidence interval. Finally we test whether the true median income difference is in the confidence interval.

Base on the Monte carol testing, our 95% confidence intercal is

```
##       2.5%      97.5%
## -3599.0826  -382.4577
```

and the true income difference is `-9000`, which is not in the confidence interval. Therefore, we can reject the null hypothesis that the median income of males and females are the same.

Based on the hypothesis tests, we find that in STEM fields, there truly exists gender base salary gap in STEM fields, which answers our first questions. So our next step is to consider whether the gender salary difference vary depending on other factors.

Firstly, we want to figure out what factors are correlation to the base salary of employees in STEM field, so we fit an linear regression model to predict base salary of employees.

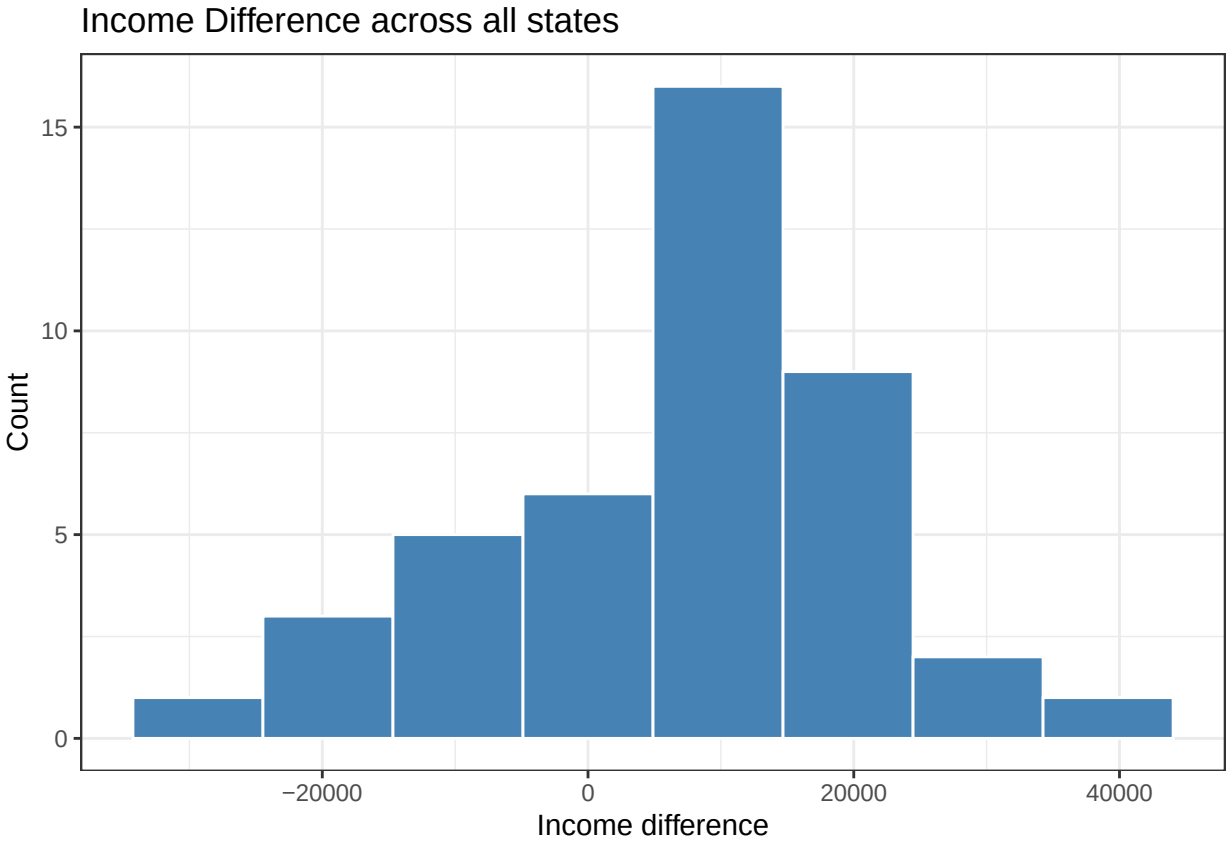
```
##
## Call:
## lm(formula = basesalary ~ 1 + gender + yearsofexperience + yearsatcompany +
##     title + costIndex + new_location, data = filted_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -283306  -16461    -615   17252 1465575
##
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       435901.21  873135.73   0.499   0.6176
## genderMale                          4424.31     652.47   6.781 1.21e-11 ***
## yearsofexperience                   4162.13      50.32  82.718  < 2e-16 ***
## yearsatcompany                     -1138.13      85.02 -13.386  < 2e-16 ***
## titleData Scientist                37280.52    2518.76  14.801  < 2e-16 ***
## titleHardware Engineer             21812.48    2554.35   8.539  < 2e-16 ***
## titleHuman Resources                5268.65    4043.10   1.303   0.1925
## titleManagement Consultant         28782.48    3056.92   9.416  < 2e-16 ***
## titleMarketing                     13366.12    3213.21   4.160 3.19e-05 ***
## titleMechanical Engineer            8730.61    3503.62   2.492   0.0127 *
## titleProduct Designer              22734.36    2645.01   8.595  < 2e-16 ***
## titleProduct Manager               28631.04    2396.05  11.949  < 2e-16 ***
## titleRecruiter                     -3789.92    3647.61  -1.039   0.2988
## titleSales                         -3268.09    3943.68  -0.829   0.4073
## titleSoftware Engineer             28797.05    2248.61  12.807  < 2e-16 ***
## titleSoftware Engineering Manager  41489.91    2483.53  16.706  < 2e-16 ***
## titleSolution Architect            16158.52    2921.16   5.532 3.20e-08 ***
## titleTechnical Program Manager     22733.34    2763.38   8.227  < 2e-16 ***
```

```
## costIndex                          -4288.71    9709.63  -0.442   0.6587
## new_locationAR                    -18756.18   30989.88  -0.605   0.5450
## new_locationAZ                     35933.11   69292.87   0.519   0.6041
## new_locationCA                    319106.64  599912.80   0.532   0.5948
## new_locationCO                     85252.59  152483.26   0.559   0.5761
## new_locationCT                    167885.72  366955.51   0.458   0.6473
## new_locationDE                     82999.72  176834.14   0.469   0.6388
## new_locationFL                     38583.85   77982.79   0.495   0.6208
## new_locationGA                      8730.02   11096.96   0.787   0.4315
## new_locationHI                    483612.12 1000429.80   0.483   0.6288
## new_locationIA                      2497.04   12333.74   0.202   0.8396
## new_locationID                     11029.70   26266.38   0.420   0.6745
## new_locationIL                     37415.26   45271.11   0.826   0.4085
## new_locationIN                      -271.66    9909.16  -0.027   0.9781
## new_locationKS                     -7833.54   14612.39  -0.536   0.5919
## new_locationKY                      3635.10   16273.25   0.223   0.8232
## new_locationLA                     13608.38   40703.91   0.334   0.7381
## new_locationMA                    203953.74  404779.00   0.504   0.6144
## new_locationMD                    189047.58  386365.79   0.489   0.6246
## new_locationME                    113108.71  269157.61   0.420   0.6743
## new_locationMI                      2875.80   13418.98   0.214   0.8303
## new_locationMN                     62956.80  113750.25   0.553   0.5799
## new_locationMO                    -16742.00   28863.46  -0.580   0.5619
## new_locationMS                    -14660.53   49461.04  -0.296   0.7669
## new_locationMT                     62345.74  166080.26   0.375   0.7074
## new_locationNC                     37189.07   49115.55   0.757   0.4490
## new_locationND                     39405.92   92212.10   0.427   0.6691
## new_locationNE                      -906.04   15908.63  -0.057   0.9546
## new_locationNH                     89807.89  192465.99   0.467   0.6408
## new_locationNJ                    177535.14  341689.89   0.520   0.6034
## new_locationNM                      5952.19   30118.92   0.198   0.8433
## new_locationNV                     71623.61  180877.99   0.396   0.6921
## new_locationNY                    260694.22  477587.31   0.546   0.5852
## new_locationOH                      8444.72   12541.24   0.673   0.5007
## new_locationOK                    -20325.11   32247.78  -0.630   0.5285
## new_locationOR                    209671.94  430021.35   0.488   0.6258
## new_locationPA                     66318.50  114700.44   0.578   0.5631
## new_locationRI                    129185.73  286598.88   0.451   0.6522
## new_locationSC                     29584.61   59587.12   0.496   0.6195
## new_locationTN                     -5616.65   15583.69  -0.360   0.7185
## new_locationTX                     22333.39   17471.10   1.278   0.2012
## new_locationUT                     47205.95   82820.00   0.570   0.5687
## new_locationVA                     68716.09  105007.48   0.654   0.5129
## new_locationVT                    106608.63  239499.01   0.445   0.6562
## new_locationWA                    125500.01  201932.18   0.621   0.5343
## new_locationWI                     46701.70   72258.26   0.646   0.5181
## new_locationWV                           NA         NA      NA       NA
## new_locationWY                     21415.20   45889.03   0.467   0.6407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44730 on 34925 degrees of freedom
## Multiple R-squared:  0.3008, Adjusted R-squared:  0.2996
## F-statistic: 234.8 on 64 and 34925 DF,  p-value: < 2.2e-16
```

According to the regression model,males have a high potential ability to earn more base salary, and some job titles, years of experiences also affect the base salary of employees. However, it seems like state location do not effects the employees, so do location effects the gender base salary gap?

Based on the regression model, we want to figure out whether the gender base salary gap very in factors liek job title, location (in state level) and companies.

First we focused on the location factors, we want to figure out whether regions (at state level) influence the income difference of males and females. We compute the income difference for each state and make the following frequency plot.



The overall histogram shows that the values range from around -20000 to near 40000 with a normal distribution, which means for many states, median income for males is 10 thousand dollars greater than females, but there are also plenty of states that have more serious income differences and some states have smaller income differences. Therefore, it seem that at the state level, there exists base salary differences of males and females.

To verity there exists base salary differences of males and females in state level, we conduct a chi-squared test to verity it. If there is no obvious gender base salary gap in state level, the stand deviation of the income difference should be very small (close to 0)

Our null hypothesis is $H_0: $ no obvious gender base salary gap in state level and H_1 : There exists gender base salary gap in state level

```
##
##  Chi-Squared Test on Variance
##
## data:  state_rela$sal_diff
## Chi-Squared = 8.5671e+13, df = 42, p-value < 2.2e-16
## alternative hypothesis: true variance is not equal to 1e-04
```

```
## 95 percent confidence interval:
##  138678637 329521430
## sample estimates:
##  variance
## 203978959
```

According to the chi-squared test, because the p-value is smaller than `2.2e-16`, which means the we can reject other null hypothesis that there is no obvious gender base salary gap in state level.

Then we want to figure out which states have gender base salary gap, and which states do not have. So for each state, we conduct a two sample wilcoxon test to test whether the median base salary of male and female in that state is equal. Then we show the test result as follows:

```
##    Row State      P_Value significant
## 1    1    CA 8.330696e-38        TRUE
## 2    2    WA 3.004607e-33        TRUE
## 3    3    MA 1.390441e-05        TRUE
## 4    4    NY 2.182609e-19        TRUE
## 5    5    SC 6.225668e-01       FALSE
## 6    6    OR 4.728207e-04        TRUE
## 7    7    VA 1.647016e-01       FALSE
## 8    8    CO 1.453442e-01       FALSE
## 9    9    NE 1.594100e-01       FALSE
## 10  10    PA 8.402782e-01       FALSE
## 11  11    IN 6.037835e-01       FALSE
## 12  12    WI 5.124009e-02       FALSE
## 13  13    TX 1.493283e-02        TRUE
## 14  14    MN 3.346728e-02        TRUE
## 15  15    IL 1.088802e-01       FALSE
## 16  16    NJ 1.555425e-01       FALSE
## 17  17    AZ 3.752096e-01       FALSE
## 18  18    NC 4.118402e-03        TRUE
## 19  19    CT 6.301570e-01       FALSE
## 20  20    NM 5.714286e-01       FALSE
## 21  21    GA 8.977133e-01       FALSE
## 22  22    FL 9.473036e-03        TRUE
## 23  23    UT 1.920187e-01       FALSE
## 24  24    AR 4.763092e-01       FALSE
## 25  25    VT 1.000000e+00       FALSE
## 26  26    IA 1.000000e+00       FALSE
## 27  27    KS 5.016907e-01       FALSE
## 28  28    MI 8.499536e-01       FALSE
## 29  29    OH 2.505470e-01       FALSE
## 30  30    NH 3.483951e-01       FALSE
## 31  31    MD 4.147890e-01       FALSE
## 32  32    TN 3.107200e-01       FALSE
## 33  33    MO 1.366325e-01       FALSE
## 34  34    DE 8.184499e-01       FALSE
## 35  35    AL 1.000000e+00       FALSE
## 36  36    ID 3.820266e-01       FALSE
## 37  37    NV 7.365355e-01       FALSE
## 38  38    KY 2.538700e-01       FALSE
## 39  39    RI 4.547610e-01       FALSE
## 40  40  <NA>           NA          NA
## 41  41  <NA>           NA          NA
```

```
## 42  42    OK 5.039434e-02       FALSE
## 43  43    ME 1.000000e+00       FALSE
## 44  44    MT 2.666667e-01       FALSE
## 45  45    MS 1.000000e+00       FALSE
## 46  46  <NA>           NA          NA
## 47  47  <NA>           NA          NA
## 48  48  <NA>           NA          NA
```

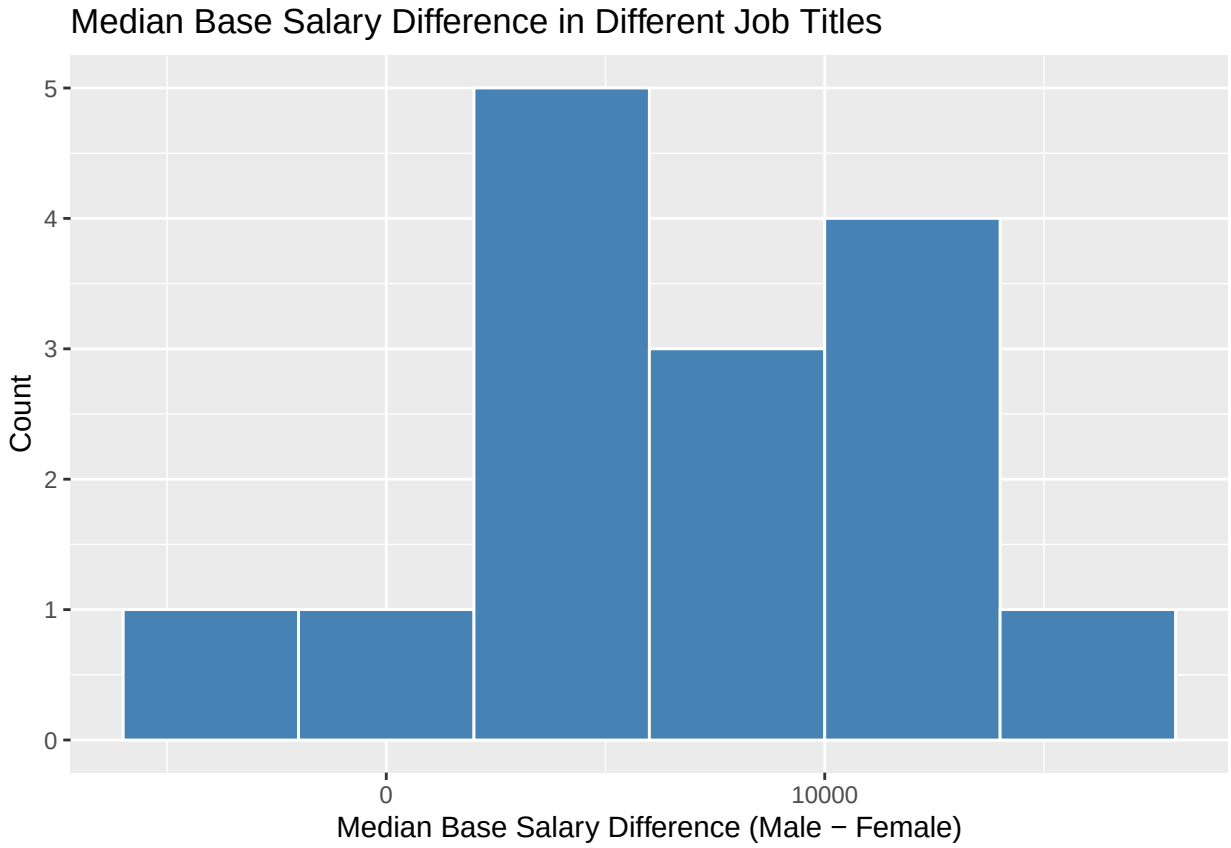
```
## # A tibble: 2 x 2
##   significant     n
##   <lgl>       <int>
## 1 FALSE          34
## 2 TRUE            9
```

According to the results of our wilcoxon test, we can find that based on our data, there are 9 states reject our null hypothesis that there are no base salary difference. And data from 34 states cannot reject the null hypothesis. Ststes that reject the null hypothesis are as follows:

```
##   Row State      P_Value significant
## 1   1    CA 8.330696e-38        TRUE
## 2   2    WA 3.004607e-33        TRUE
## 3   3    MA 1.390441e-05        TRUE
## 4   4    NY 2.182609e-19        TRUE
## 5   6    OR 4.728207e-04        TRUE
## 6  13    TX 1.493283e-02        TRUE
## 7  14    MN 3.346728e-02        TRUE
## 8  18    NC 4.118402e-03        TRUE
## 9  22    FL 9.473036e-03        TRUE
```

Therefore, at least based on our data and test results, there is obvious evidence that CA, WA, MA, NY, OR, TX, MN, NC, and FL states has gender base salary gaps, which means the the base salary difference in gender are vary related to location.

Is there any difference in gender gaps for different job positions?



From this histogram, we can see that for most job titles, male workers have a higher median base salary than female workers. For some job titles like Data Scientist, Recruiter, the difference is very small. For Management Consultant, Human resources, and Sales, the difference is relatively large. Therefore, it seems that, within each job title, there exists base salary differences of males and females.

To verity that base salary differences of males and females in each job title is not the same, we conduct a chi-squared test to verity it. If there is no obvious gender base salary gap in each job level, the standard deviation of the base salary difference should be very small. And we consider the a standard deviation less than 500 to be small.

Our null hypothesis is

$H_0: $ No obvious gender base salary gap difference between each job title and H_1 : There exists gender base salary gap difference between each job title

```
##
##  Chi-Squared Test on Variance
##
## data:  gap$basesalary
## Chi-Squared = 3.805e+12, df = 14, p-value < 2.2e-16
## alternative hypothesis: true variance is not equal to 1e-04
## 95 percent confidence interval:
##  14567968 67599665
## sample estimates:
## variance
## 27178571
```

According to the chi-squared test, because the p-value is smaller than `2.2e-16`, which means the we can reject the null hypothesis that there is no obvious gender base salary gap difference between each job title.

Then we want to figure out how big the gender gap in base salary for different job titles. So for each job title, we conduct a two sample wilcoxon test to test whether the median base salary of male and female with that job title is equal. Then we show the test result as follows:

```
##    Row                     Job_Title      P_Value significant
## 1    1             Business Analyst 1.833379e-01       FALSE
## 2    2               Data Scientist 1.294007e-01       FALSE
## 3    3            Hardware Engineer 3.711855e-02       FALSE
## 4    4              Human Resources 5.569707e-02       FALSE
## 5    5        Management Consultant 1.796622e-02       FALSE
## 6    6                    Marketing 1.501460e-02       FALSE
## 7    7          Mechanical Engineer 7.165900e-01       FALSE
## 8    8             Product Designer 1.347460e-05        TRUE
## 9    9              Product Manager 2.052063e-13        TRUE
## 10  10                    Recruiter 5.036747e-01       FALSE
## 11  11                        Sales 5.107261e-01       FALSE
## 12  12            Software Engineer 2.458608e-26        TRUE
## 13  13 Software Engineering Manager 8.542828e-02       FALSE
## 14  14           Solution Architect 7.946010e-03        TRUE
## 15  15    Technical Program Manager 1.802876e-03        TRUE
```

According to the results of our wilcoxon test, we find that we can reject the null hypothesis (Male workers have the same median base salary as female workers with the same job title) for 5 job titles. For the rest 10 job titles, we cannot reject the null hypothesis.

```
##   Row                 Job_Title      P_Value significant
## 1   8          Product Designer 1.347460e-05        TRUE
## 2   9           Product Manager 2.052063e-13        TRUE
## 3  12         Software Engineer 2.458608e-26        TRUE
## 4  14        Solution Architect 7.946010e-03        TRUE
## 5  15 Technical Program Manager 1.802876e-03        TRUE
```

In conclusion, gender gap presents differently in different job titles and in some jobs, there might not be a gender gap. From the above histogram and wilcoxon test, we find that there is strong evidence that gender gap exists in employees with job title “Product Designer”, “Product Manager”, “Software Engineer”, “Solution Architect”, and “Technical Program Manager”.

Summary

Known Problems

Possible Future Questions