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 base salary 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:

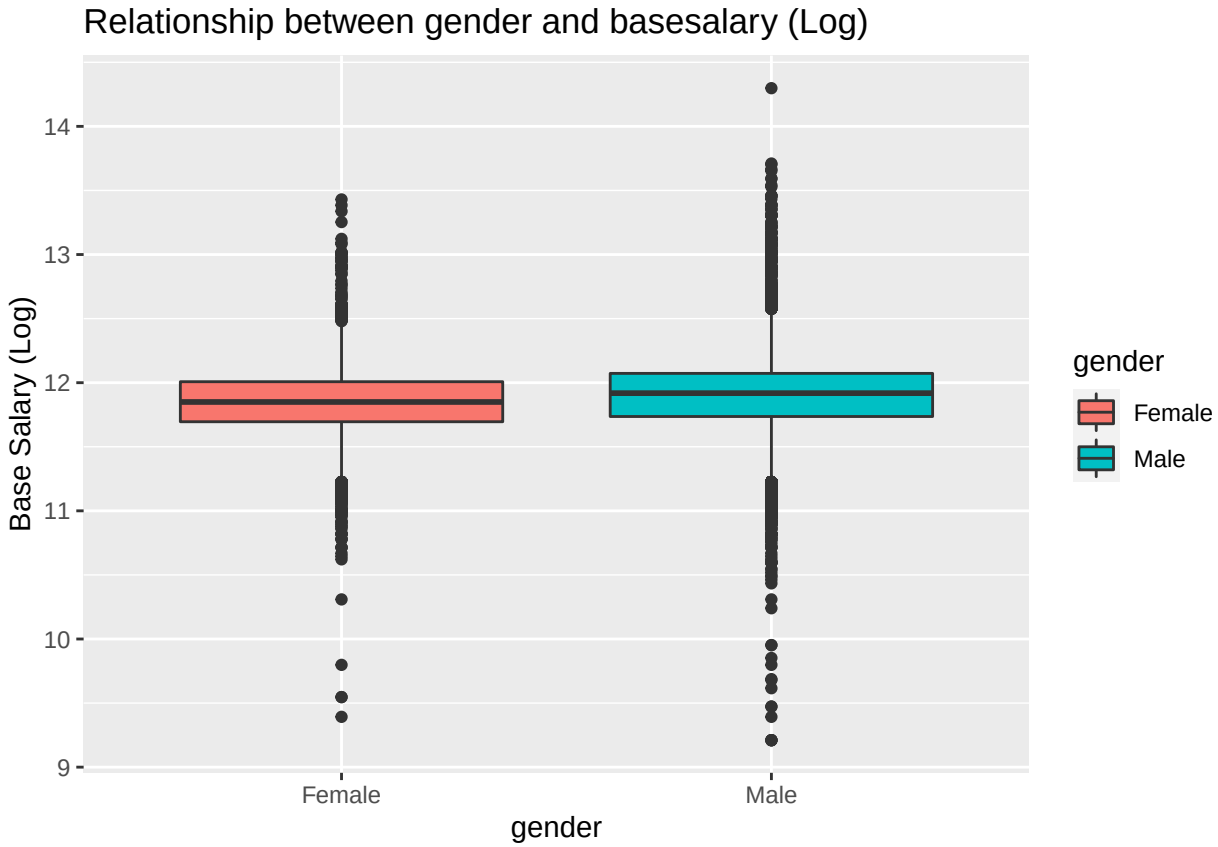
Main question: Is there a gender gap in income level(base salary) in different data science and STEM subfields and different regions? Does the difference vary depending on other factors (e.g., education, subfields, companies, regions, etc.)?

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 'basesalary' (in logarithm) of males and females. We use logarithm to better visualize the difference

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



According to the above plot, It seems that the median base salary of females is only slightly lower than males base salary. Therefore, we cannot tell whether males and females truly have different median base salary 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 base salary for males and females are not the same.

We also apply the Monte carol testing to test the null hypothesis. Firstly, we adapt the permutation testing approach by randomly assign the base salary data into male group and female groups, and compare their medians, store the median base salary 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 base salary difference is in the confidence interval.

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

```
##       2.5%      97.5%
## -3639.6061  -315.9269
```

and the true base salary difference is `-9000`, which is not in the confidence interval. Therefore, we can reject the null hypothesis that the median base salary 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.

Throwing every state into the model as a predictor is not a great approach, here. We do variable selection rather than just throwing them all into a model. Certainly many of these states will have an effect– CA, WA, NY and MA are very expensive states and many tech companies are based there, so salaries will be higher (both due to cost of living and to having more management roles).Therefore, we only include some states as as predictors and include cost of living data about different statesin linear regression model.

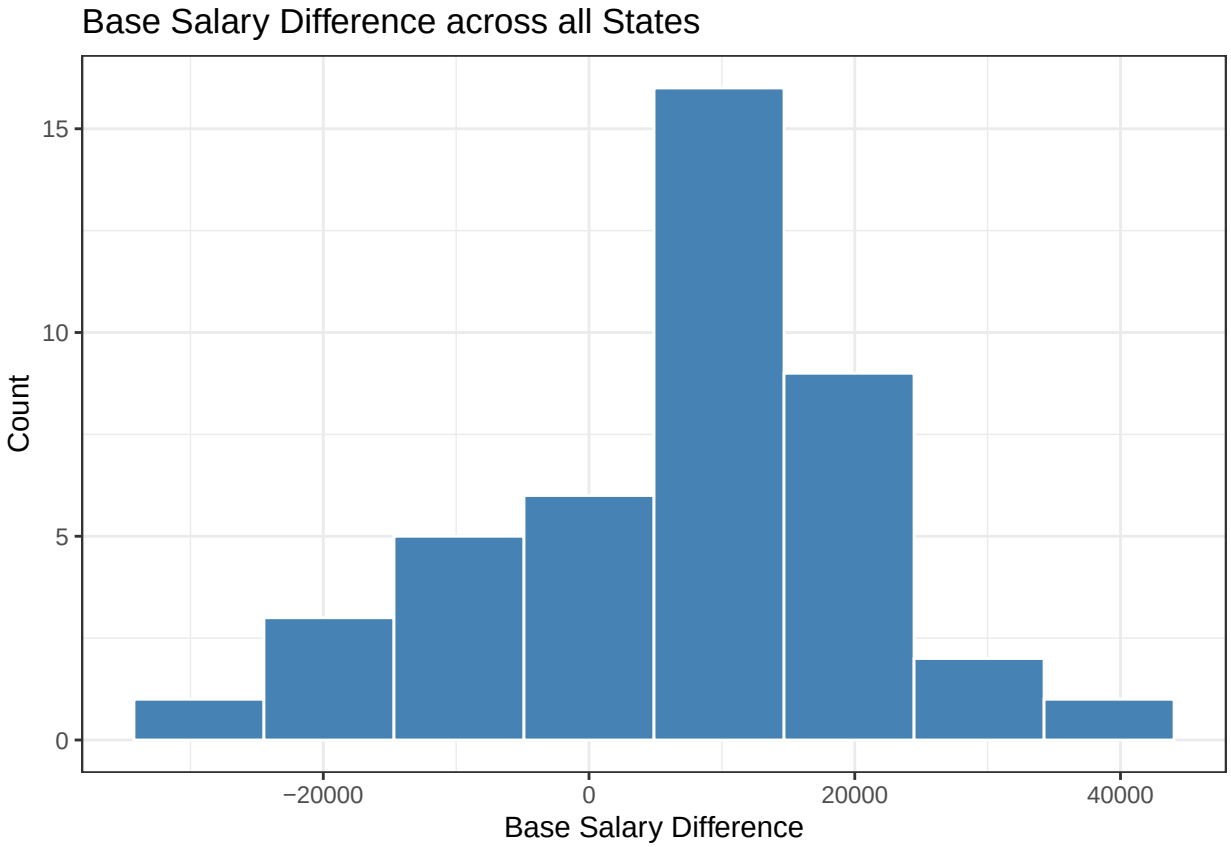
```
##
## Call:
## lm(formula = basesalary ~ 1 + gender + yearsofexperience + yearsatcompany +
##     title + costIndex + select_location, data = filted_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -283438  -16751    -520   17545 1465795
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      68076.34    6580.81  10.345  < 2e-16 ***
## genderMale                        4393.94     653.07   6.728 1.75e-11 ***
## yearsofexperience                 4164.22      50.35  82.705  < 2e-16 ***
## yearsatcompany                   -1172.31      84.91 -13.806  < 2e-16 ***
## titleData Scientist              37401.70    2518.52  14.851  < 2e-16 ***
## titleHardware Engineer           21948.06    2547.58   8.615  < 2e-16 ***
## titleHuman Resources              5517.14    4044.36   1.364   0.1725
## titleManagement Consultant       29899.03    3051.51   9.798  < 2e-16 ***
## titleMarketing                   13612.14    3214.42   4.235 2.29e-05 ***
## titleMechanical Engineer          7527.91    3501.60   2.150   0.0316 *
## titleProduct Designer            22879.43    2643.71   8.654  < 2e-16 ***
```

```
## titleProduct Manager               28902.10    2394.42  12.071  < 2e-16 ***
## titleRecruiter                     -3448.79    3649.02  -0.945   0.3446
## titleSales                         -2299.17    3945.52  -0.583   0.5601
## titleSoftware Engineer             28990.88    2246.74  12.904  < 2e-16 ***
## titleSoftware Engineering Manager  41936.64    2482.35  16.894  < 2e-16 ***
## titleSolution Architect            16420.17    2920.32   5.623 1.89e-08 ***
## titleTechnical Program Manager     23118.43    2762.72   8.368  < 2e-16 ***
## costIndex                            238.82      40.69   5.869 4.42e-09 ***
## select_locationMA                 -24172.87    1576.99 -15.329  < 2e-16 ***
## select_locationNY                  -1418.84    1015.19  -1.398   0.1622
## select_locationOther              -27458.47    2213.71 -12.404  < 2e-16 ***
## select_locationWA                  -7984.16    1777.23  -4.492 7.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44800 on 34967 degrees of freedom
## Multiple R-squared:  0.2978, Adjusted R-squared:  0.2974
## F-statistic: 674.1 on 22 and 34967 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. Also, it seems like state location indeed effects the employees' salary (some state location is significant to the base salary, and cost of living also make the contribution of the base salary), 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 vary in factors like job titles, locations (in state level) and companies.

First we focused on the location factors, we want to figure out whether regions (at state level) influence the base salary difference of males and females. We compute the base salary 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 base salary for males is 10 thousand dollars greater than females, but there are also plenty of states that have more serious base salary differences and some states have smaller base salary differences. Therefore, it seem that at the state level, there exists base salary differences of males and females.

To verify there exists base salary differences of males and females in state level, we conduct a chi-squared test to verify it. If there is no obvious gender base salary gap in state level, the stand deviation of the base salary 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>
```

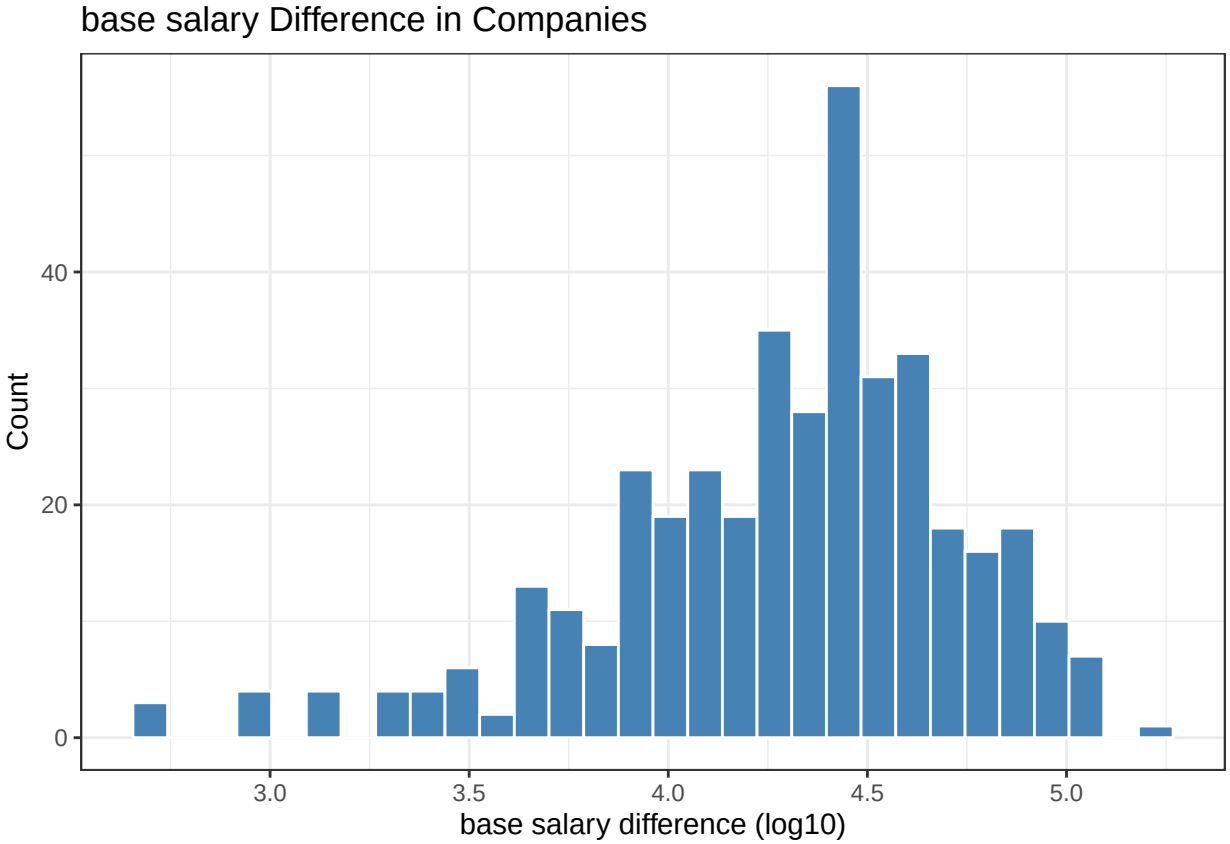
```
## 1 FALSE          34
## 2 TRUE            9
```

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.

Now we focused on the companies factors, we want to find out which companies have the most influence on the base salary difference of males and females. The plot follows the same logic as above-we count the number of companies which has base salary difference according to gender.



At this level, we can observe there are significant base salary differences between males and females in most companies. Next step we will test how many companies are in the confidence interval.

```
##     company              Male            Female          sal_diff
```

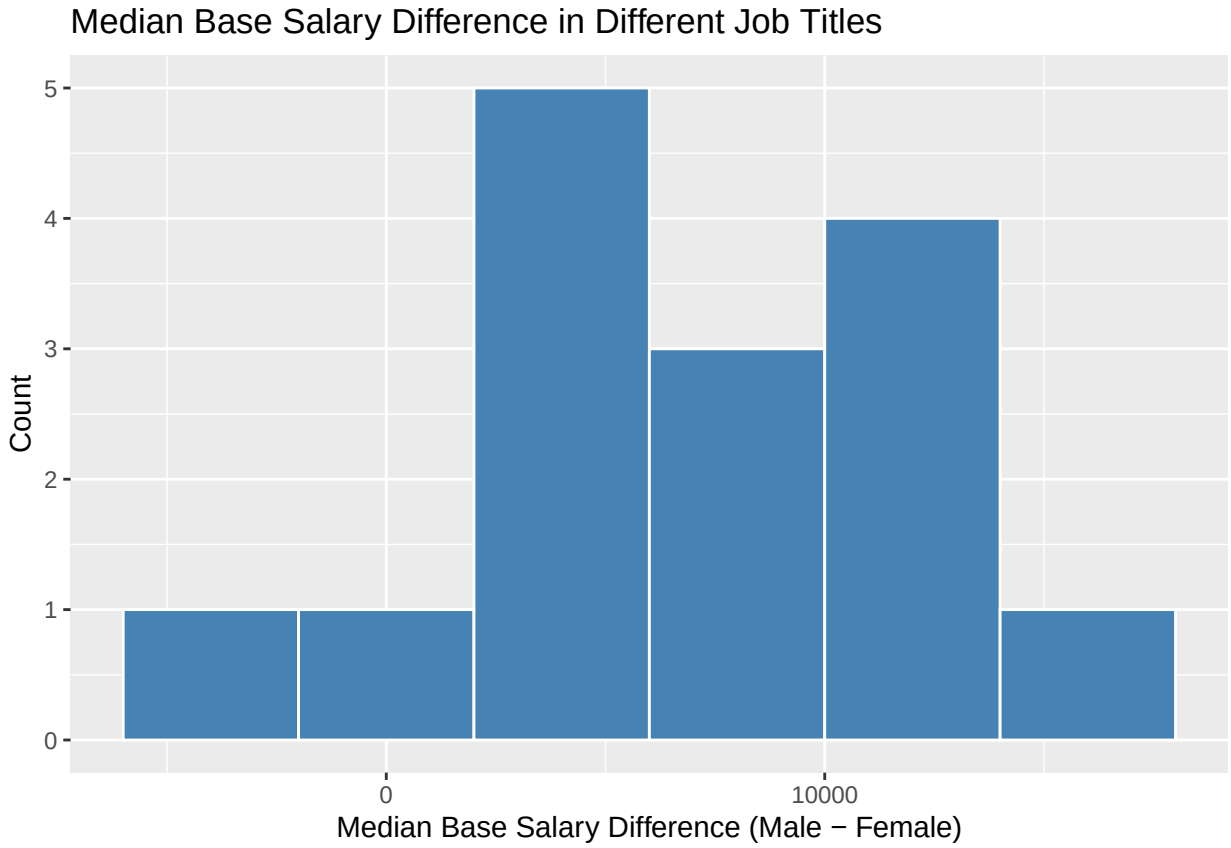
```
##  Length:614         Min.   :     0   Min.   :     0   Min.   :-174000
##  Class :character   1st Qu.:116625   1st Qu.:103000   1st Qu.:  -7375
##  Mode  :character   Median :138000   Median :129000   Median :  10000
##                     Mean   :139286   Mean   :128192   Mean   :  11094
##                     3rd Qu.:161000   3rd Qu.:150000   3rd Qu.:  28500
##                     Max.   :455000   Max.   :405000   Max.   : 167500
##        x           Sign_diff
##  Min.   :   2.00   Mode :logical
##  1st Qu.:   5.00   FALSE:568
##  Median :   9.00   TRUE :46
##  Mean   :  54.35
##  3rd Qu.:  27.00
##  Max.   :4612.00
```

First, we assume that the base salary difference between men and women is zero in all companies, which means that the base salary difference is closer to zero, the more fair the base salary between men and women is. Next, we let zero be central to the confidence interval because we want this confidence interval to capture the companies that have smaller salary differences of gender.

```
##             company    x Sign_diff
## 20           AMAZON 4612     FALSE
## 349      MICROSOFT 3101     FALSE
## 239         GOOGLE 2574     FALSE
## 201       FACEBOOK 1872     FALSE
## 32           APPLE 1309     FALSE
## 397         ORACLE  660     FALSE
## 273          INTEL  622     FALSE
## 463     SALESFORCE  597     FALSE
## 114          CISCO  549     FALSE
## 262            IBM  531      TRUE
## 558           UBER  485     FALSE
## 95     CAPITAL ONE  437     FALSE
## 315       LINKEDIN  428     FALSE
## 579         VMWARE  373      TRUE
## 76       BLOOMBERG  331     FALSE
## 434       QUALCOMM  330     FALSE
## 295 JPMORGAN CHASE  305     FALSE
## 277          INTUIT  287     FALSE
## 409          PAYPAL  255      TRUE
## 585        WAYFAIR  248     FALSE
```

As result, we can see most companies are not captured by a confidence interval("FALSE" in "Sign_diff"), which means they have gender bias salaries. Only ten percent of companies are "TRUE" for "Sign_diff",which means they have fair salaries.

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

The alternative hypothesis is $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

Based on the hypothesis tests, we find that in STEM fields, there exists gender base salary gap in STEM fields. So next we wanted to determine whether the gender base salary difference vary depending on other factors. Based on the regression model, we figured out that the gender base salary gap very may have relationship in factors like job title, location (in state level) and companies. After we used chi-squared test and wilcoxon test, the result shows that gender gap exists in employees with job title "Product Designer", "Product Manager", "Software Engineer", "Solution Architect", and "Technical Program Manager". And we let zero be central to the confidence interval to capture the companies that have smaller salary differences of gender. We found most companies have gender bias salaries, only "IBM", "VMWARE" and "PAYPAL", or 10% of companies have fair salaries. Finally, according to the histogram, the result of the wilconxon and the chi-squared test, there are 9 states reject our null hypothesis that there are no base salary difference, they are CA, WA, MA, NY, OR, TX, MN, NC, and FL, which means they have gender base salary gaps. In Conclusion, the base salary difference in gender are vary related to job title, companies and location.

Known Problems

Our data has 29 columns, and each data entry may have NA value in different columns. If we simply drop all the rows containing NA value would make our dataset super small. So we decided to only drop the rows with

NA value in the 'gender' column, and for other columns, we will have some extra data cleaning before using them for different purposes.

Another issue is that there are some outliers in our dataset (some people with extremely high salary), including their data in our dataset might lead to a worse performance for linear regression models and other models.

Since There are a lot more males than females in the data set (female sample is 2/3 of the male sample), it would influence the accuracy of our estimated model. Data imbalance is not only a problem in classification task, but also in regression tasks. The performance of a regression model may suffer from the fact that the distribution of the target variable is not normally distributed and skewed. Applying transformations on the target variable can boost the performance.

Possible Future Questions

So far, we can can only determine that there exist a gender gap in income level(base salary) in different data science and STEM subfields and different regions and the difference vary depending on different location(states level), different data science and STEM subfields and different companies. However, we can not qualify the gender gap in base salary, and we need to deal with solving the problem of sample imbalance in our dataset. To be able to deal with imbalanced data using these models, you have one of two options: first, is to increase the representation of the observations of interest vs. the other observations (or vice versa). Second, is to adapt the model itself by parameter tuning based on customized criteria. Also, it is worth further explore how the salary difference by other possible factors in STEM fields as well as non-STEM fields.

Reference

Carlton, G. (2021, February 22). The biggest barriers for women in STEM: BestColleges. BestColleges.com. Retrieved November 12, 2021, from https://www.bestcolleges.com/blog/barriers-for-women-in-stem/.

Stephen J. Ceci, Donna K. Ginther, Shulamit Kahn, Wendy M. Williams. (2015, January 1). Do Women Earn Less Than Men in STEM Fields? Scientific American. Retrieved December 04, 2021, from https://www.scientificamerican.com/article/do-women-earn-less-than-men-in-stem-fields/