Exploring Salary

Christy Nelson & Jorge Rodriguez

Comp 4447 – Fall 2023

Dataset and Motivation:

When we started the ideation process to see what would it be a good project to do all the things we learned in Data Science Tools COMP 4447, we found several references to a very well defined dataset with several key columns that will allowed to be use in many directions; from just simple data exploration all way to Machine Learning; so we thought that this dataset will allow us to put in practice everything we learned like dataframes creations, cleaning data, data wrangling, data visualization and correlation. The dataset was found on Kaggle and contains information on the following characteristics of 6702 employees:

* Age
* Gender
* Education Level
* Job Title
* Years of Experience
* Salary

Actual Task Definition/Research Question:

We will examine the relationship between the different aspects of the employees and see how they contribute to or seem to be related to salary, including any group trends, with a particular focus on salary vs gender.

Literature Review

One existing study in particular that exists shows a statistically significant gender gap in hiring for men vs. women showing in a completely randomized experiment that male application materials were given a higher starting salary offer and hire ability rating than identical application materials with a female name.

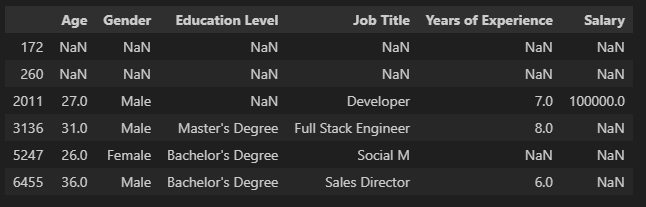
* Original study: Moss-Racusin, C., Dovidio, J., et al. “Science faculty’s subtle gender biases favor male students.” PNAS October 9, 2012 109 (41) 16474-16479; https://doi.org/10.1073/pnas.1211286109

Many of the jobs in this dataset are related to STEM fields. Below are two more studies showing a gap in STEM fields that favor men vs women.

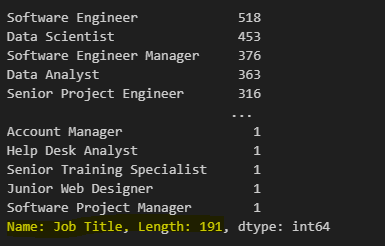
* -Huang, J. et al. “Historical comparison of gender inequality in scientific careers across countries and disciplines.” *Proceedings of the National Academy of Sciences,* Mar 2020, 117 (9) 4609-4616; DOI: 10.1073/pnas.1914221117
* -*Boston Consulting Group,* “What’s Keeping Women out of Data Science?” bcg.com/publications/2020/what-keeps- women-out-data-science.aspx

Data Cleaning & Outliers:

After obtaining the dataset, we examined the data column by column and row by row, starting by checking for NaN values and we found several records that have missing information on them, further analysis of each record, we agreed that none of them made sense to fill out with a mean or other value, so we proceed to remove all rows containing NaNs.



After cleaning the NaN records, we continued by standardizing the differences in education level names. For example, there were multiple expressions of the same name level such as “Master’s Degree” and “Master’s. We also found an education level named “other”, which after further analysis of this category, we decided we needed to be removed from the Gender category because it was too small (14 values) to serve as any good comparison.



Then we looked at the number of job titles, which was nearly 200, including several ones with only 1 employee on it, so we created a new smaller dataframe that eliminated all jobs with fewer than 40 workers on it to better analyze the data, the relationships, and correlations in addition to streamline displays and remove any anecdotal information.

After reducing amount of Job Titles with less than 40 employees on it, we ended up with a good 44 titles to further or research, so we moved into the outlier’s identification, it made the most sense to look for stay data in regards to the response variable (salary). If we look at the dataset as a whole, there are no individual outliers, which would be indicated with a separate point above or below the boxplot:

A blue rectangular object with black lines

Description automatically generated

Salary Boxplot

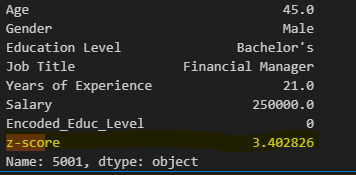
However, if we look at all the subgroups by gender and education level; for example, there are outliers at every level as shown below:

A diagram of a group of boxes

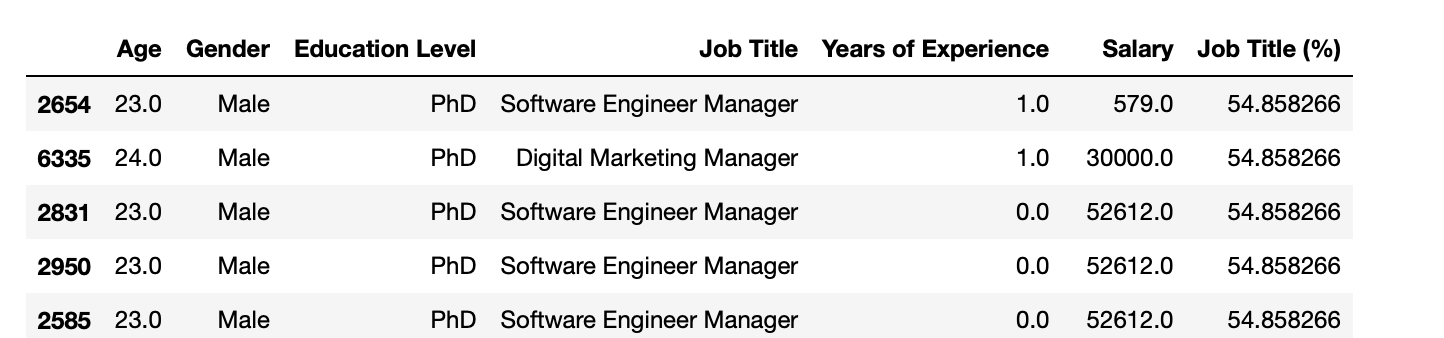
Description automatically generated with medium confidence

When we looked at removing these outliers, starting with the outlier in the upper left corner, which is the highest overall salary and the only outlier at the bachelor’s level, this only reduces the overall male average salary by only $22.00 and the next highest male salary is only $10,000 lower than that point. Because of this minor effect, it seems like any subgroup outliers do not need to be removed from the data as disruptions but rather stand out as overachievers in their subgroup according to salary. Taking a closer look at this observation, for example, we see a male employee working as a Financial Manager with 21 years of experience. It makes sense that someone with this job and this level of experience would be able to earn a higher salary than the “typical” employee with a bachelor’s degree.

In addition, we also numerically calculated outliers of the overall dataset and created upper and lower arrays of outliers, which were both empty. However, when we added a column of z-scores grouped by education and salary, this Financial Manager was confirmed an outlier by having a z-score of 3.4.



We also looked at the lower end of the PhD employees:



And found that, while not calculated to be an outlier of the overall set, observation #2654 is probably an inaccurate data value at a salary of $579/year, so we feel it would be justified to remove him.

Visualization:

Data Summary as a whole:

The dataset contains more male employees:

A blue and orange rectangular bars

Description automatically generated

The majority of the employees have a Bachelor’s Degree, with a progressively smaller number of employees having a Master’s and then a PhD and the smallest number having only a high school education:

A graph of a bar chart

Description automatically generated with medium confidence

We were able to identify the top ten paying jobs, which you can see are all in the STEM field:

A graph with different colored lines

Description automatically generated with medium confidence

1. **Is there a relationship between education level and gender?**

Here is a cross-tabulation of Education Level vs Gender:

A table with numbers and text

Description automatically generated

We created a barplot to visualize this data:

A graph of a graph with blue and orange bars

Description automatically generated

After looking at the graph, it does not seem to be any overall pattern with gender at any level of education – it alternates between fewer men limited to a high school education to more men with a bachelor’s degree then back to fewer men with a master’s and more men with a PhD. Therefore, there is not a corresponding trend with one gender dominating when it comes to having more education than the other.

This is confirmed by the low Cramer’s V value between these variables of 0.19.

1. **Is there a relationship between Salary and Gender?**

We have just seen above that education level does not seem to depend on gender. However, if we look at salary by education level instead of just number of people, you can see that male employees have a higher salary at every education level, even if there are fewer member of the dataset of that gender.

A graph of a graph of a graph

Description automatically generated with medium confidence

Here is a closer examination of these categories, to revisit the graph we looked at when examining outliers:

A diagram of a graph

Description automatically generated with medium confidence

We can see that nearly in every marker for females in every group, the salary is lower than the corresponding marker for males. Which is to say that the minimum salary for females is lower in each education level, except for the PhD level, where the only salaries below the female’ minimum salaries are the outliers, or atypical data; the same pattern as the first quartile, median, third quartile and maximum.

In addition to the finding above, if we zoom out and look at the data set as a whole split by gender, we see the same trend of lower quartiles for women overall.

A graph showing a bar chart

Description automatically generated with medium confidence

1. **Do these gender differences pervade across career titles?**

Below we can see a barplot of the job titles most represented in this dataset, and the Percentage of each job represented by males vs females. We can clearly see that 51.08% of the Jobs are Male Vs. 48.91% of Female jobs, so we can see a gap of ~2.00% of the Job Market with the majority being dominated by Males.

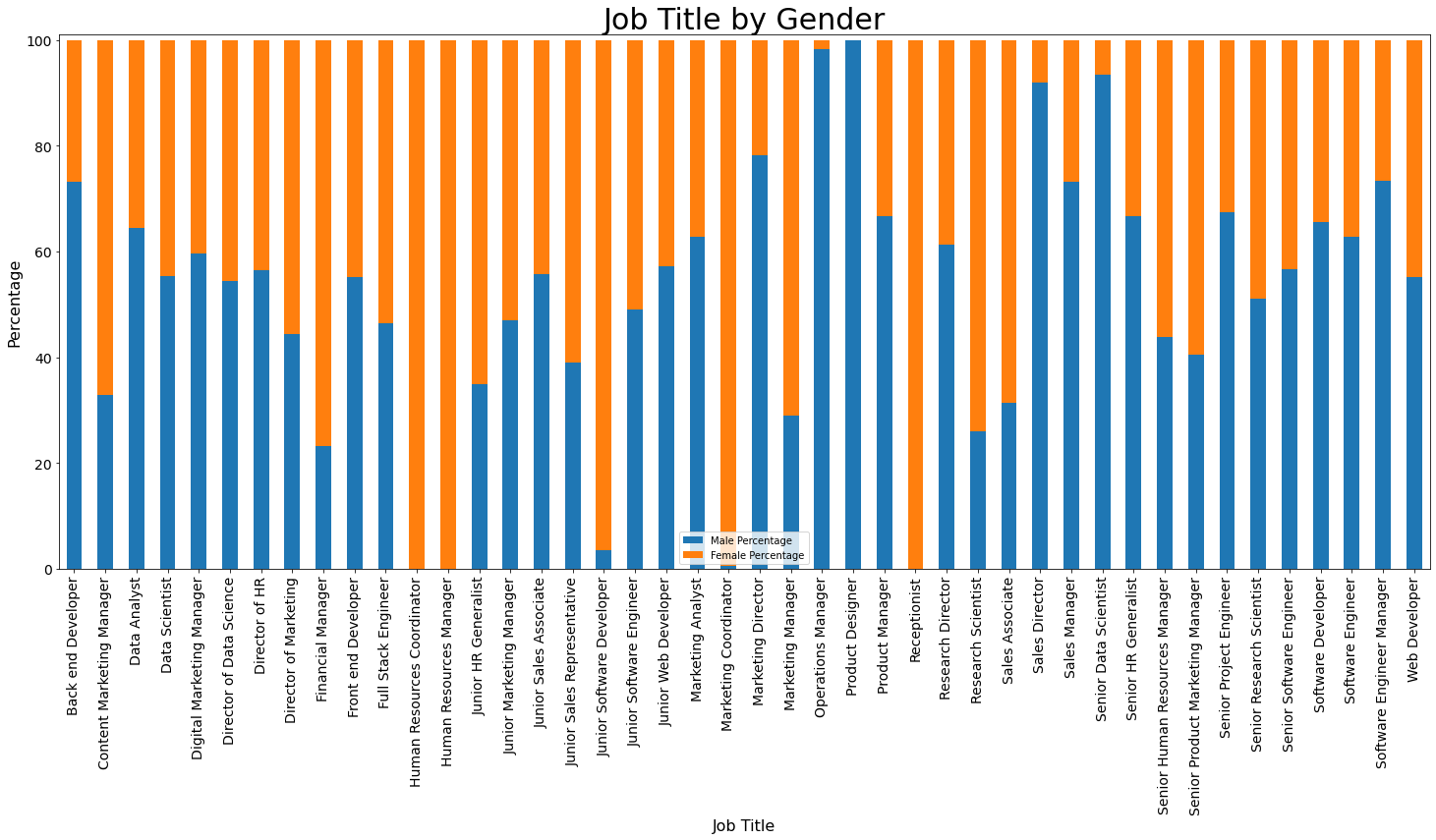


Because we know the count of males is overall higher than the count of females, we felt it was important to look at the breakdown of percentages instead of pure count. This will also allow us to look at which jobs are dominated by which gender.

For example, you can see from the stacked bar graph below that there are some jobs solely represented by females (human resources jobs and receptionist); however none of them are in the STEM field. Job fields that contain both genders but have a larger percentage of female workers include: Content Marketing Manager, Director of Marketing, Financial Manager, Junior HR Generalist, Junior Marketing Manager, Junior Sales Associate, Junior Software Developer, Marketing Manager, Research Scientist, Sales Associate, Senior HR Manager, and Senior Product Marketing Manager. Of those, at best 2 can be considered STEM fields. Thus you can see that fields being dominated by women are largely avoidant of STEM. Additionally, women are the majority of 5 of the 7 jobs with the word “Junior” and only 2 of the 7 with the word “Senior.” This makes an interesting statement on their ability to advance in the workplace and their ability to lead vs the ability of men.

Contrarily, men make up the larger portion of more jobs containing the title “Senior”, “Director, and “Manager”, such as Software Engineer Manager, Sales Manager, Sales Director, Research Director, Product Manager, Director of HR, and Director of Data Science.

Additionally, they dominate more jobs in the STEM fields, including Data Analyst, Data Scientist, Front End Developer, Back End Developer, Software Developer, Software Engineer, and Web Developer.

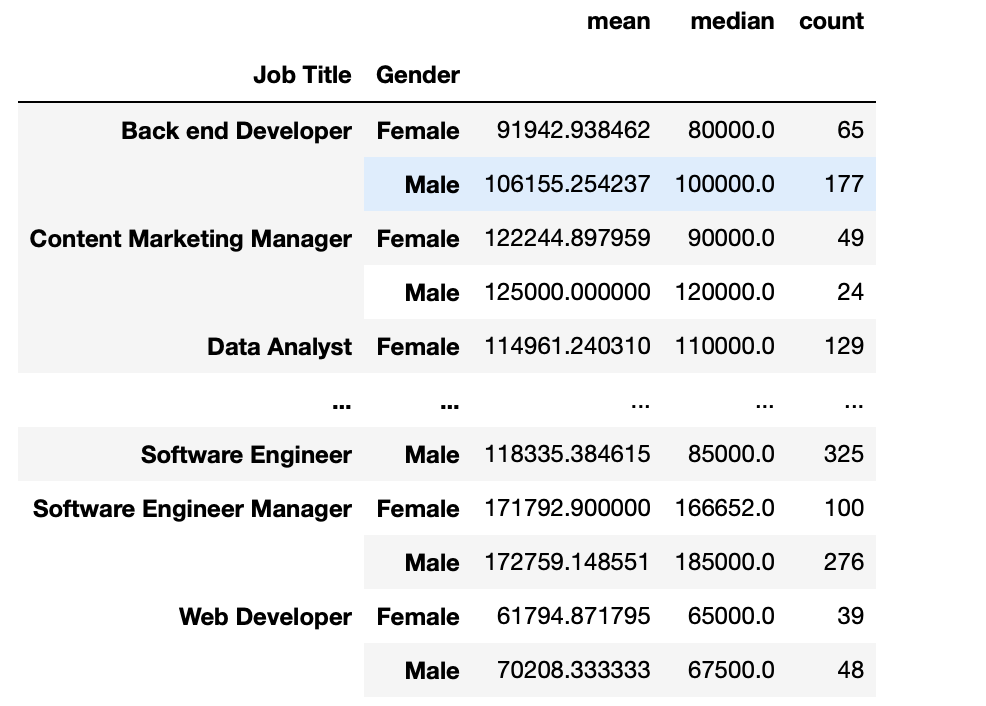


When we graph by job title and salary by gender, 28 out of the 40 jobs (70%) of the jobs offer a higher average salary to men than to women.

A graph of different colored lines

Description automatically generated

The exact amounts can be accessed by viewing this table:



1. **What are the relationships between the quantitative variables of this dataset?**

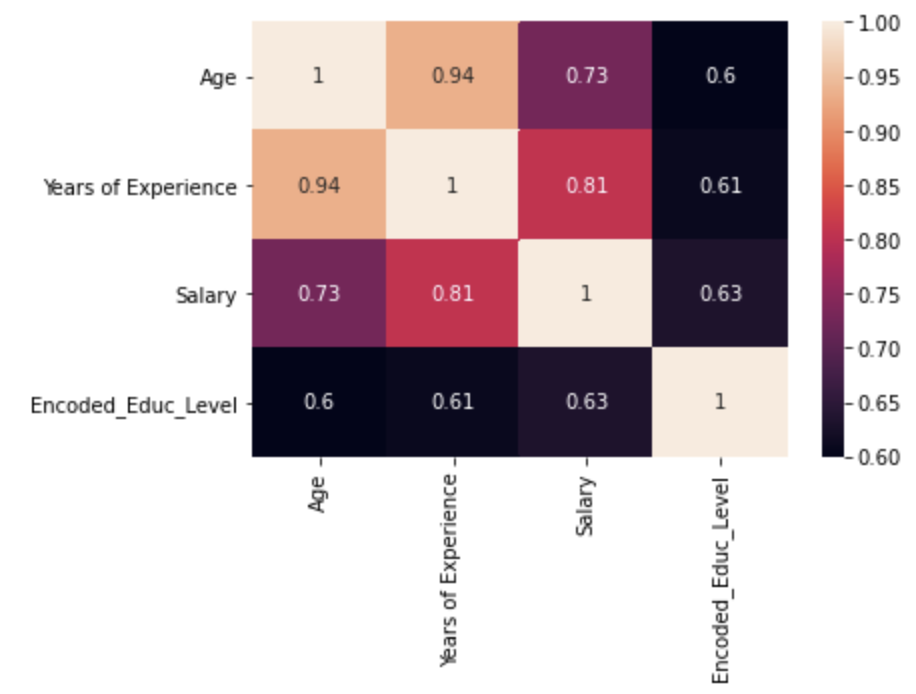
A graph showing a number of dots

Description automatically generated with medium confidence

A graph of a chart

Description automatically generated with medium confidence

As expected, you can see from these two scatterplots that salary increases with both age and years of experience. This trend is true for both men and women. We can further examine the strength of these relationships by looking at a heatmap of correlations between all variables, including the properly re-encoded (with 0 = High School, 1 = Bachelor’s, 2 = Master’s, 3 = PhD) Education Level:



The strongest positive linear relationship is between age and years of experience with a value of 0.94, followed by salary and years of experience with a value of 0.84. The relationship between salary and age is still extremely strong (though a little weaker) at 0.73, and Encoded Education Level has a positive linear relationship of similar strength with all of age, salary, and years of experience.

If we look at the overall statistics of the dataset without regards to subdivisions, we see that women make a lower salary, on average, by about $14,000/year but also have about a year and a half less experience on average.

A number on a white background

Description automatically generated

A graph of a person with blue and orange squares

Description automatically generated A graph of a salary per gender

Description automatically generated

Throughout all the data and visualizations, we have shown, this pattern is consistent with the relationships we saw in the scatterplots, barplots and heatmaps, perhaps women should have a lower average salary if they have a lower average age and a lower average number of years of experience. This makes sense until we look at the violin plot of years of experience according to gender, where we see that even though there are more men in the data set, there are more women with more years of experience. One possible explanation of this is that since they made less per year than their male counterparts, they need to work longer in order to accumulate the same retirement or nest egg.

A blue and orange shapes

Description automatically generated

Conclusions:

This data set seems to confirm both a gender gap in STEM fields as well as a salary gap between women and men overall. Because this dataset consists of over 6000 employees across 200 career titles, it seems that the differences are significant and pervasive. However, since this is an observational study and not a controlled experiment, we can at best say there is a correlation and not that gender causes the gap.