The Effects of Gender on Salary and Workforce Composition by Industry

Ryan Verhoef

Author Note

This paper is intended to demonstrate my data analysis capabilities. The code used to conduct the analysis can be found in the accompanying Jupyter Notebook.

Abstract

This paper explores the gender pay gap and its connection to industry representation in the United States. The analysis utilizes a dataset obtained from a survey conducted on askamanager.org, focusing on respondents' demographics, job-related information, and income. The paper aims to investigate the relationship between gender representation and wages across different industries, shedding light on the disparities that exist in the workforce. The dataset, consisting of 27,966 entries, underwent rigorous data cleaning and preparation to ensure reliability. Filtering out records with missing or inconsistent values, as well as excluding entries from countries other than the United States, resulted in a final dataset of 19,939 records representing 20 industries. The analysis reveals a moderate effect of industry on gender distribution, indicating uneven representation across various sectors. The analysis also highlights that increasing gender representation alone may not guarantee equal pay. The correlation between gender representation and the ratio of pay was found to be weak, indicating the influence of additional factors.

Keywords: Gender pay gap, gender representation in the workforce, gender equality.

The Effects of Gender on Salary and Workforce Composition by Industry

Introduction

Gender inequality in the workplace, particularly in terms of salary disparities, has long been a topic of concern and research. The primary reason women make less is because they tend to work in lower paying industries. However, this is not the only factor driving low wages which is demonstrated by a gender pay gap within most industries (Foster et al., 2020; Palffy et al., 2022). This paper aims to investigate the relationship between gender representation in different industries and the corresponding wage disparities as well as reproduce previous research on a new dataset that I found on askamanager.org (Green, 2021). While most research compares women and men, I wanted to include respondents who identified as non-binary in the analysis. Since non-binary individuals comprise less than 3% of the dataset that I analyzed and there were not enough of them to compare their salaries to the rest of the dataset within each industry. I therefore classified non-binary individuals as non-males along with women since they are both considered minority genders.

Methods

I downloaded the dataset used in this analysis from the blog askamanager.org (Green, 2021). The blog collected the data from a survey hosted on the site. The survey asked respondents questions about their age, the industry they're employed in, their job title, if there is any additional context to their job title, their annual salary, how much additional compensation the receive if any, the currency their paid in, if there is any additional context to their income, what country they work in, what state they work in of they work in the US, how many years of professional work experience they have both in their field and overall, highest level of education, gender, and race. Readers could access the data collected from the survey on a Google spreadsheet linked through the site. The dataset contains 27,966 total entries.

Data Cleaning, Preparation and Validation

Since I focused the research question on the US, I filtered out records that had an empty value for the state column. I also filtered out records where the respondent entered a different country than the United States such as Belgium or Zimbabwe despite selecting a US state. One respondent indicated that they lived in the US but received payments in Japanese Yen from Japanese companies for translation work through the internet while living in the US. I filtered out this record was since they were working for a company entirely outside of the US and may have distinct cultural norms around wages. Even after I excluded those records, there were still 22 records where the respondents had indicated that they were paid in a currency other than US dollars. Looking at the records, these respondents still entered cities that are in the states they indicated and there was nothing else that indicated that the respondents worked outside of the US. I included these records in the analysis because I assumed that they miskeyed their entry when selecting currency.

There were six entries in the gender column of the dataset: “man,” “woman,” “non-binary,” “prefer not to answer,” “Other or prefer not to answer,” or an empty value. Since the original research question is focused on comparing males to minority gender, I classified respondents that entered “man” as “Male” and those that entered either “woman” or “non-binary” as “non-male.” I excluded answers where the respondent preferred not to answer or where there was an empty value since there was no way to tell what their actual gender may be.

Data Preparation

I aggregated the data by industry and gender. I excluded industries that had fewer than 30 records in either category of gender from the data since a sample of fewer than 30 for a given combination of industry and gender could potentially skew the results due to small sample size. After filtering out those industries with too few respondents of either gender, there were 19,939 records in 20 different industries that I will use in the analysis. 3,908 of those respondents identified as male and 16,031 identified as either female or non-binary (see Table 1).

Validation

The dataset as is heavily right skewed is evident in the histogram in Figure 1. The boxplot of the data in Figure 1 shows that there are many outliers past the rightmost 1.5 inter-quartile range represented by horizontal lines corresponding to around the $200,000 point on the plot. Upon investigation of some of the records that were outliers there was no clear pattern as to why many respondents reported unusually high salaries. For some of the entries for additional context to salary, some comments indicate that their reported salary figure includes stock options, bonuses, and other supplementary compensation despite not reporting it separately on the additional compensation question. The question in the survey also instructs part time and hourly workers to annualize their compensation by multiplying their hourly rate by 40 hours for the week and 52 weeks for the year. Some of the entries in the additional context question indicate that many respondents followed this instruction despite working less than 40 hours per week or 52 weeks per year.

To reduce the potential impact of outliers in the analysis, the salary data were Winsorized[1]. There was not enough evidence of systematic data entry error to justify trimming the outliers, but there were enough concerns with their validity as mentioned previously to justify limiting their effect on the mean to ensure that they do not impact the final analysis. Winsorization has the added benefit of not affecting the analysis of the proportion of the genders since the process modifies rather than eliminates records in the dataset. As a final precaution, I ran two analyses in parallel: one with the outliers unaltered and the other with the Winsorized dataset. This will allow us to see how the outliers might affect the analysis (see Table 2 and Figure 2).

Analysis

I measured the dependance of the gender composition of the workforce on industry using Pearson's chi squared test. I conducted the test using the chi squared contingency function from the stats module of the SciPy Python package. The results showed a moderate effect using Cramér's V which

demonstrates unequal distributions of genders in different industries X^2 (20, N = 19,939) = 1713.12, p < .001, Φ_c = .29 (see Figure 3).

I compared the mean salary of males compared to non-males using Welch's *t*-test using the stats module from SciPy. I evaluated both the Winsorized and non-Winsorized data for statistically significant differences to examine and compare the effects of Winsorization. Prior to analysis, the statistical significance threshold to accept the alternative hypothesis (α) that I established at .05. In both sets of data, art and design, education (primary/secondary), entertainment, hospitality & events, marketing, advertising and PR, retail, and transport or logistics did not have a *p*-value less than .05. In the non-Winsorized data, I also found that business or consulting, property or construction, and sales did not meet the predefined α (see Table 3-4).

I calculated the effect size for the statistically significant industries and all industries combined for each set of data using Cohen's *d* for samples with unequal variance (Cohen, 1988) (see Table 3-4). In the context of this dataset, we will consider a *d* less than or equal to 0.3 a small effect size and a *d* greater than 0.3 but less than or equal to 0.7 will be considered to have a moderate effect size (see Table 3-4 and Figure 4)

For each industry analyzed, I calculated the Pearson correlation coefficient (*r*) comparing the gender ratio with overall mean salary (see Figure 5) and the ratio of male salaries to non-male salaries (see Figure 6). The correlation between gender ratio and mean overall salary is moderate for both the non-Winsorized data r (18) = .52, p < .05 and the Winsorized data r (18) = .53, p < .05. The correlation between gender ratio and ratio of male salaries to non-male salaries was weak and failed to reject the hypothesis that there is no relationship between representation in the industry and equal pay for both the non-Winsorized data r (18) = -.003, p = .98 and the Winsorized data r (18) = -.21, p = .35.

Discussion

The results of the chi square test show that there is a moderately uneven distribution of genders across the various industries analyzed in this sample. This could in turn be a factor in lower overall wages for minority genders with the difference in male salaries in the Winsorized dataset (M = 113,242, SD = 55,584) compared to non-male salaries (M = 86,263, SD = 43,542), t (19,938) = 28.3, $p < 0.001$, d = 0.54. Indeed, the correlation between salaries and the total mean salary showed that minority genders are more likely to work in lower paying industries. This is consistent with previous research showing that differences in representation is one of the leading causes for the overall lower wages of women and minority genders (Foster et al., 2020; Palffy et al., 2022). However, there was no statistically significant correlation found between the ratio of genders and the ratio of pay. This would suggest that increasing minority gender representation in each industry with lower levels of representation may not lead to increased pay. This also shows that while unequal gender representation across industries is a contributing factor in overall wage inequality, it cannot be the only factor driving the gender pay gap.

The analysis of the gender pay gap by industry found that the difference in mean pay was most pronounced in industries that are typically classified as either finance or tech. Insurance, accounting, banking and finance, and computing and tech, and utilities and telecommunications consistently had moderate effect sizes in both analyses. The industries that did not have a statistically significant difference in mean salary were ones that required interpersonal communication skills. These industries included art and design, primary and secondary education, entertainment, hospitality and events, marketing, advertising and PR, and retail. The one other industry that failed to have a statistically significant difference in pay that does not require interpersonal skills is transportation and logistics. This could be explained by the fact that many truckers are paid a transparent flat rate per mile regardless of whether they are an employee or an owner-operator.

Limitations

Based on the dataset and the way the data were collected, there are some concerns that the data analyzed may not be representative of the population of workers. I excluded some respondents from the analysis because they did not work in industries that had a large enough sample size to have a sufficiently powered analysis of the gender pay gap within the industry. In the final dataset used in the analysis, 77.9% of respondents identified as female, whereas the U.S. Bureau of Labor Statistics (2023) reported that women comprised only 46.8% of the total labor force.

Since the survey was only published at askamanager.org (Green, 2021), the sample may be biased to readers of that blog. The question asking about the participant's salary was worded in such a way that respondents may not have accurately reported their annual salary. The question asked respondents who worked for an hourly rate to multiply their hourly rate by 40 hours and then 52 weeks rather than the actual number of hours and weeks the respondent worked in a year. This could have biased mean salaries upwards since respondents who only worked part-time or for part of the year may have reported higher wages than their actual annual earnings. The data may also be biased by the fact that salaries were self-reported, and respondents may be tempted to overestimate how much they earn in a year.

Further Research

Since analysis of the data also only examines the relationship between variables within the dataset and does not prove causality, further analysis is necessary to understand the causes of the gender pay gap within industries or why minority genders are more likely to work in certain industries rather than others. Factors that may be at play could include socialization and norms that pressure individuals to select gender appropriate occupations. The subcultures within certain industries could also encourage gender discrimination and may be higher in some industries compared to others. Women also have a greater expectation to perform family caregiving tasks in the home, which may lead women to select occupations with greater flexibility that offer lower wages.

Conclusion

This paper investigated the relationship between gender representation and wages across different industries in the United States. The analysis of a dataset obtained from a survey conducted on askamanager.org shed light on the gender pay gap and its connection to industry distribution. The findings reveal notable disparities in wages between males and non-males, as well as unequal gender representation across various sectors.

Consistent with Foster et al. (2020), the results of the analysis demonstrate a moderate effect of industry on gender distribution, indicating that certain industries have a higher concentration of one gender over the other. This uneven representation can contribute to the overall wage inequality observed between males and non-males. Industries categorized as finance and tech consistently exhibited significant differences in mean salaries, highlighting the existence of gender-based pay gaps in these sectors. However, it is important to note that increasing gender representation alone may not necessarily lead to equal pay. The analysis did not find a statistically significant correlation between the ratio of genders in an industry and the ratio of pay, suggesting that other factors may be at play in driving the gender pay gap. Additional research is required to explore the underlying causes of this phenomenon, including socialization, occupational norms, and cultural biases that may influence occupational choices and wage disparities.

By raising awareness of the gender pay gap and its connection to industry distribution, this research encourages stakeholders, including employers, policymakers, and advocacy groups, to work collaboratively towards fostering a more inclusive and equitable workforce. Achieving gender equality in terms of wages and opportunities is not only a matter of social justice but also essential for economic growth and societal well-being. Continued research and proactive measures are vital in driving meaningful change and creating workplaces that provide equal rewards for equal work, regardless of gender.

References

Cohen, J. (1988). Chapter 2. In Statistical Power Analysis for the behavioral sciences (pp. 43–44). essay, L. Erlbaum Associates.

Foster, T. B., DeWolf, M., Murray-Close, M., & Landivar, L. C. (2020). An Evaluation of the Gender Wage Gap Using Linked Survey and Administrative Data. United States Census Bureau. https://doi.org/Working Paper Number CES-20-34

Green, A. (2021, April 27). How much money do you make?. Ask a Manager. https://www.askamanager.org/2021/04/how-much-money-do-you-make-4.html

Palffy, P., Lehnert, P., & Backes-Gellner, U. (2022). Social norms and gendered occupational choices of men and women: Time to turn the tide? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4281259

U.S. Bureau of Labor Statistics. (2023, January 25). Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. U.S. Bureau of Labor Statistics. https://www.bls.gov/cps/cpsaat11.htm

Footnotes

[1]I Winsorized the data by using Tukey's fence method and defining the lower fence as the first quartile minus the inter-quartile range (IQR) and the upper fence as the third quartile plus 2.5 times the IQR. I set the upper fence higher because I only wanted to reduce the effect of far out values, since there are always significant outliers when it comes to income.

Tables

Table 1

Number of Respondents and Mean Salary by Industry

Industry	Male Respondents	Mean Male Salary	Non-Male Respondents	Mean Male Salary
Accounting, Banking & Finance	238	$114,761	1,188	$87,514
Art & Design	43	$80,090	251	$86,554
Business or Consulting	109	$125,059	571	$101,974
Computing or Tech	1,511	$144,114	2,116	$119,461
Education (Higher Education)	240	$78,006	1,785	$67,375
Education (Primary/Secondary)	65	$66,722	638	$64,185
Engineering or Manufacturing	384	$104,587	1,000	$95,332
Entertainment	42	$111,881	157	$102,538
Government and Public Admin.	183	$92,840	1,196	$82,340
Health Care	156	$103,977	1,435	$91,255
Hospitality & Events	36	$68,327	175	$67,428
Insurance	75	$102,687	371	$81,604
Law	104	$163,030	842	$117,444
Marketing, Advertising & PR	111	$95,418	790	$89,420
Media & Digital	113	$99,497	489	$83,247
Nonprofits	180	$83,474	1,911	$72,280
Property or Construction	56	$82,799	252	$71,198
Retail	69	$70,402	319	$60,019
Sales	61	$100,056	177	$104,909
Transport or Logistics	61	$83,223	174	$76,885
Utilities & Telecommunications	71	$103,832	194	$87,010
Total	3,908	$116,504	16,031	$88,274

Note. The above table shows mean salary for the non-Winsorized data and number of respondents for each industry and gender. The totals number of respondents and mean salary for the whole data set is at the bottom of the table.

Table 2

Winsorized Mean Salary Data

Industry	Mean Male Salary	Mean Non-Male Salary
Accounting, Banking & Finance	$110,552	$86,266
Art & Design	$80,090	$75,092
Business or Consulting	$115,453	$101,404
Computing or Tech	$139,869	$117,309
Education (Higher Education)	$77,637	$66,732
Education (Primary/Secondary)	$66,744	$63,281
Engineering or Manufacturing	$103,900	$94,267
Entertainment	$110,190	$93,847
Government and Public Administration	$92,855	$81,440
Health care	$101,001	$87,907
Hospitality & Events	$68,327	$64,325
Insurance	$102,647	$80,604
Law	$134,992	$112,942
Marketing, Advertising & PR	$95,431	$88,128
Media & Digital	$99,129	$82,717
Nonprofits	$82,076	$71,416
Property or Construction	$82,799	$70,936
Retail	$69,436	$60,028
Sales	$98,039	$75,148
Transport or Logistics	$83,223	$76,893
Utilities & Telecommunications	$103,832	$87,017
Total	$113,242	$86,263

Note. The above table shows mean salaries after Winsorization. The totals number of respondents and mean salary for the whole data set is at the bottom of the table.

Table 3

t-Statistic, p-Value, and Cohen's d for Non-Winsorized Mean Salaries

Industry	*t*-statistic	*p*-value	Cohen's *d*
Insurance	3.55	0.001	0.46
Utilities & Telecommunications	3.01	0.003	0.42
Accounting, Banking & Finance	5.175	<0.001	0.41
Total	20.774	<0.001	0.38
Computing or Tech	10.166	<0.001	0.35
Law	2.621	0.01	0.34
Media & Digital	3.105	0.002	0.34
Education (Higher Education)	3.633	<0.001	0.27
Government and Public Administration	3.293	0.001	0.24
Nonprofits	2.935	0.004	0.24
Engineering or Manufacturing	3.398	0.001	0.2
Health care	2.157	0.032	0.17
Art & Design	-0.473	0.636	N/A
Business or Consulting	1.973	0.051	N/A
Education (Primary/Secondary)	0.442	0.66	N/A
Entertainment	0.692	0.491	N/A
Hospitality & Events	0.118	0.907	N/A
Marketing, Advertising & PR	1.318	0.19	N/A
Property or Construction	1.977	0.052	N/A
Retail	1.429	0.157	N/A
Sales	-0.166	0.868	N/A
Transport or Logistics	1.007	0.317	N/A

Note. I did not calculate Cohen's *d* for industries that had a *p*-value greater than .05.

Table 4

t-Statistic, p-Value, and Cohen's d for Winsorized Mean Salaries

Industry	*t*-statistic	*p*-value	Cohen's *d*
Total	28.299	<0.001	0.54
Insurance	3.8	<0.001	0.52
Accounting, Banking & Finance	6.354	<0.001	0.49
Sales	3.036	0.003	0.48
Computing or Tech	13.178	<0.001	0.45
Utilities & Telecommunications	3.009	0.003	0.42
Media & Digital	3.245	0.001	0.37
Law	3.26	0.001	0.34
Government and Public Administration	3.79	<0.001	0.32
Property or Construction	2.035	0.046	0.32
Education (Higher Education)	3.948	<0.001	0.31
Business or Consulting	2.725	0.007	0.29
Nonprofits	3.468	0.001	0.29
Engineering or Manufacturing	4.106	<0.001	0.25
Health care	2.749	0.007	0.25
Art & Design	0.704	0.484	N/A
Education (Primary/Secondary)	0.612	0.542	N/A
Entertainment	1.423	0.16	N/A
Hospitality & Events	0.617	0.54	N/A
Marketing, Advertising & PR	1.656	0.1	N/A
Retail	1.375	0.173	N/A
Transport or Logistics	1.006	0.317	N/A

Note. I did not calculate Cohen's *d* for industries that had a *p*-value greater than .05.

Figures

Figure 1

Distributions of Non-Winsorized Data

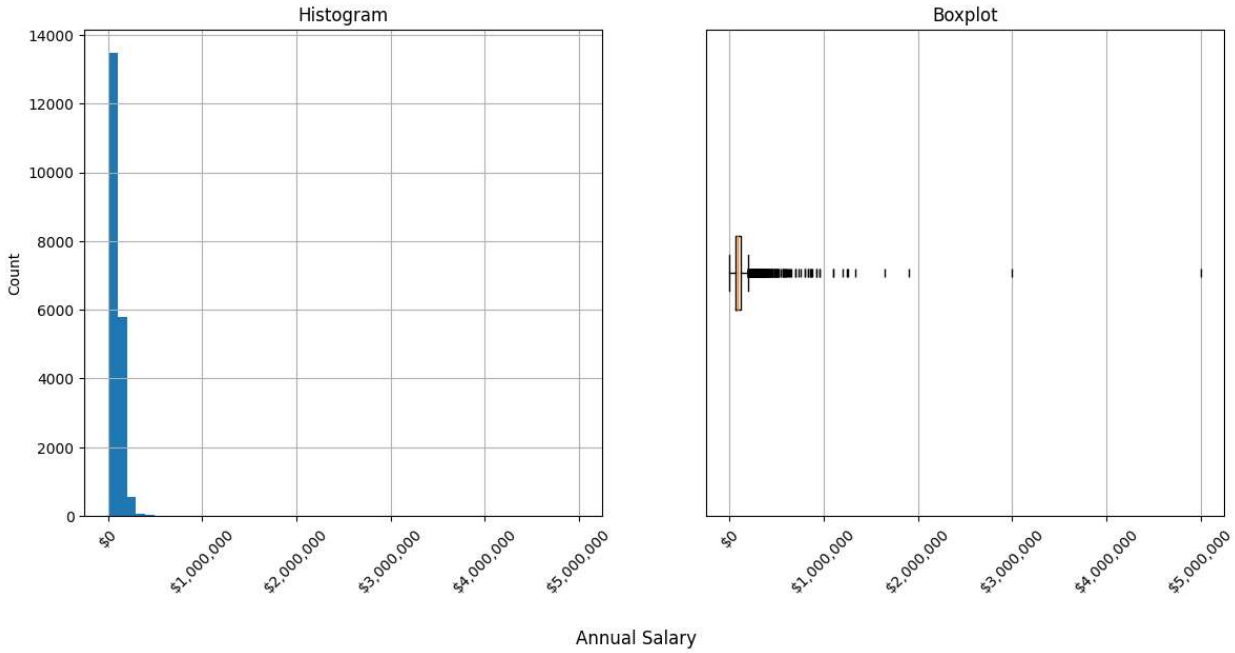


Figure 2

Distributions of Winsorized Data

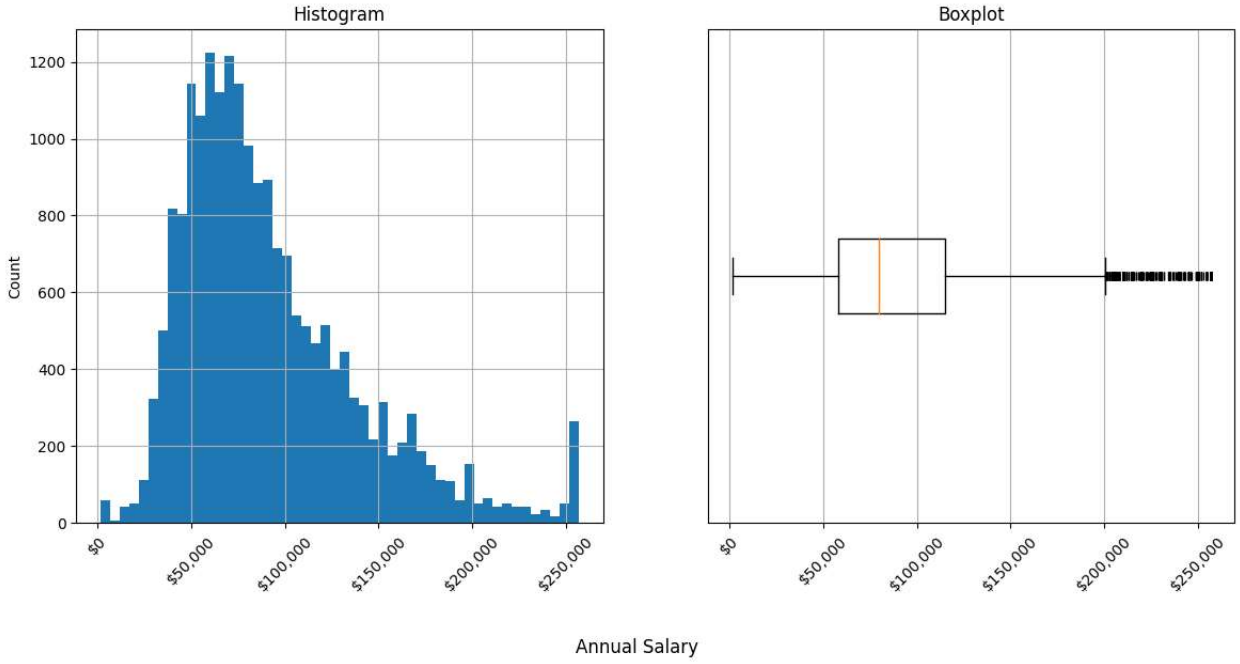
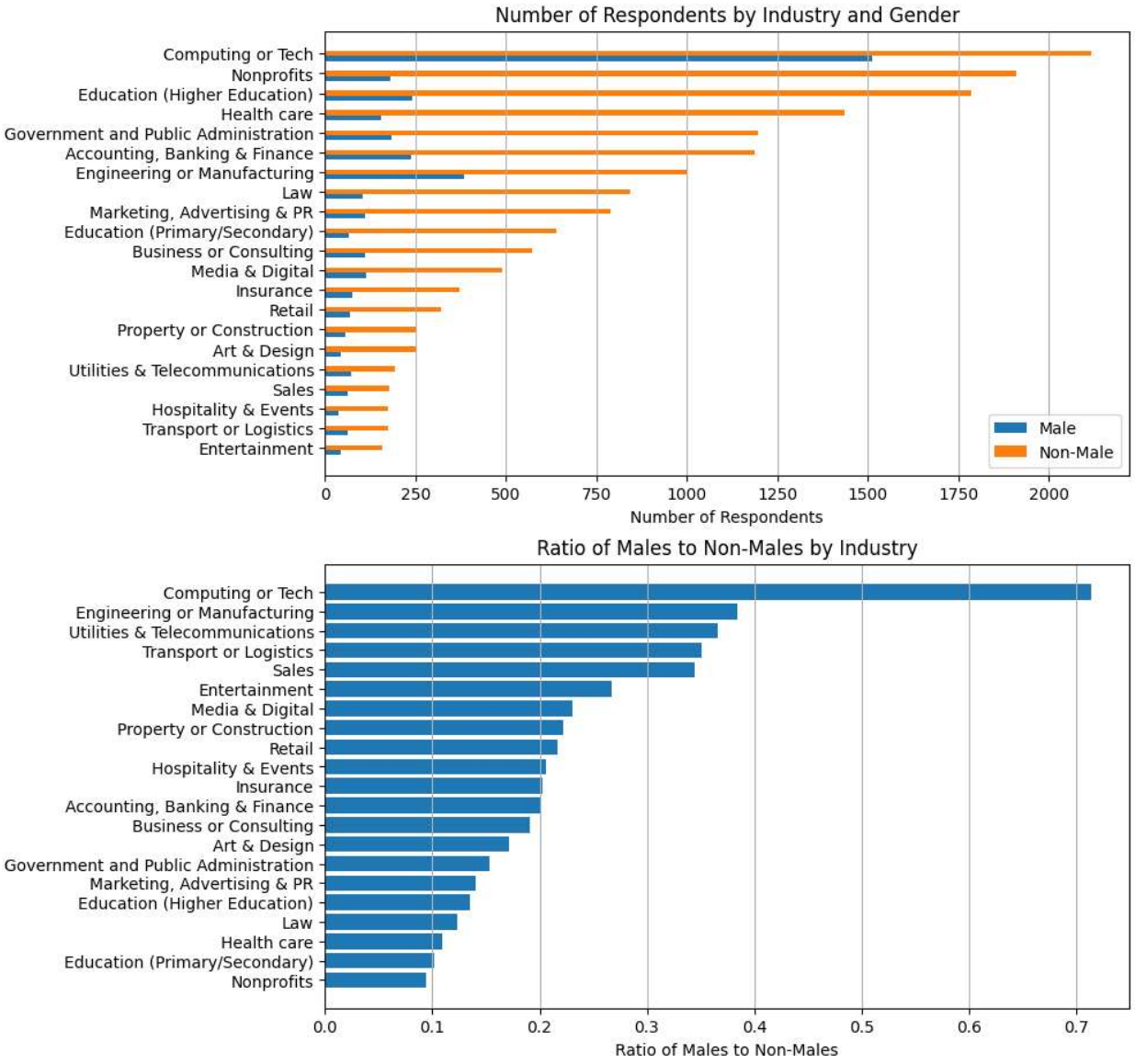


Figure 3

Number of Male vs. Non-Male Respondents



Note: Industry has a moderate effect on predicting industry based on sex with X^2 (20, N = 19,939) = 1713.12, $p < .001$, $\Phi_c = .29$.

Figure 4

Average Salary by Industry and Gender

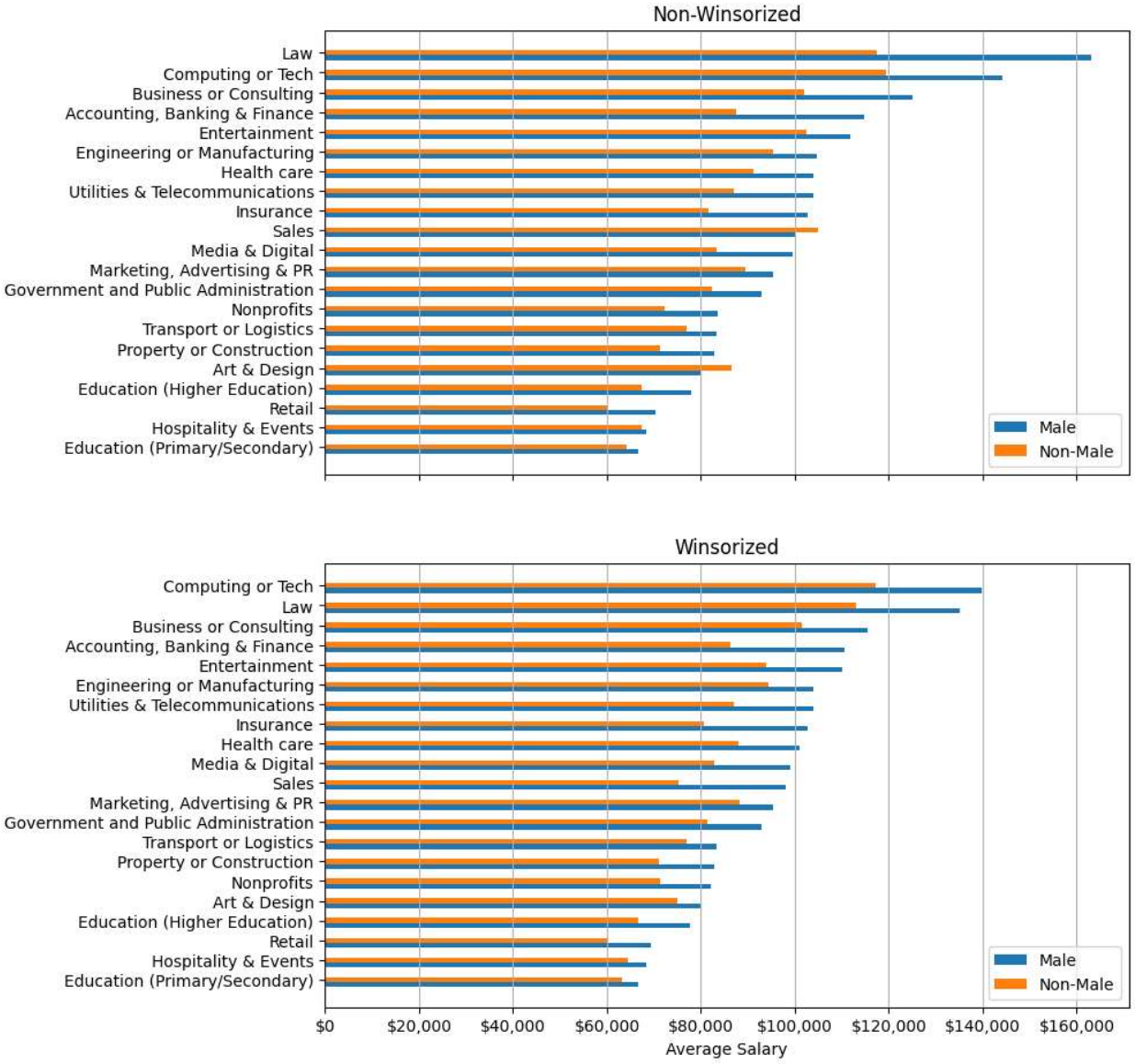
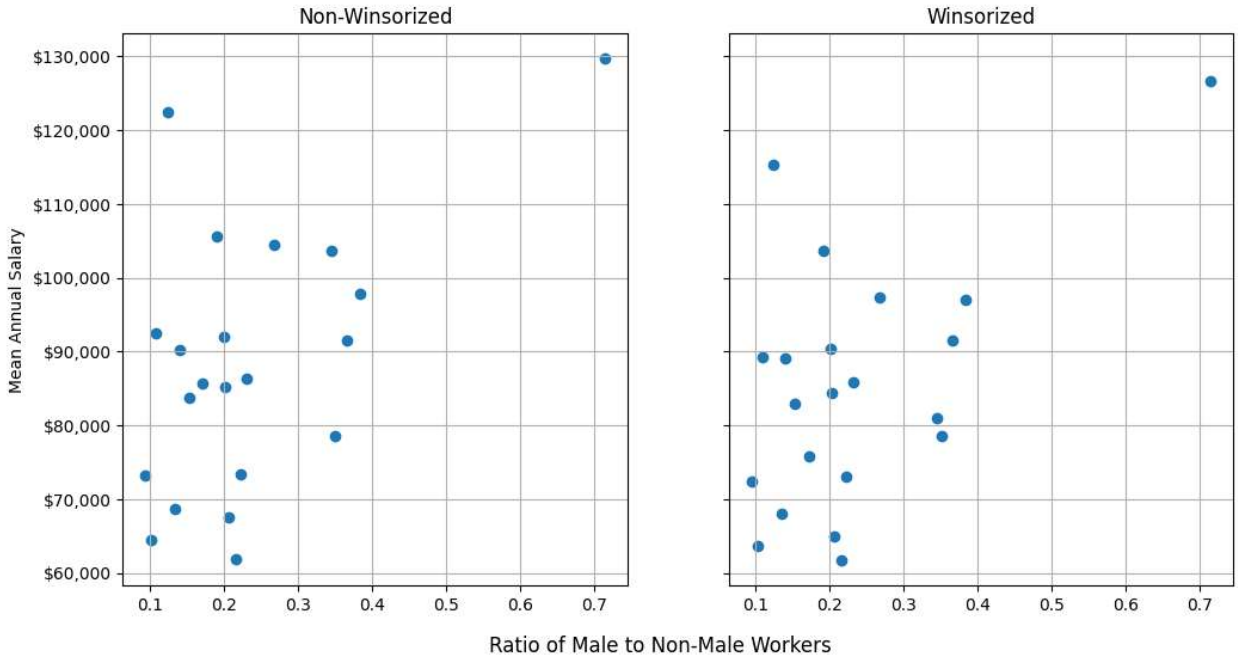


Figure 5

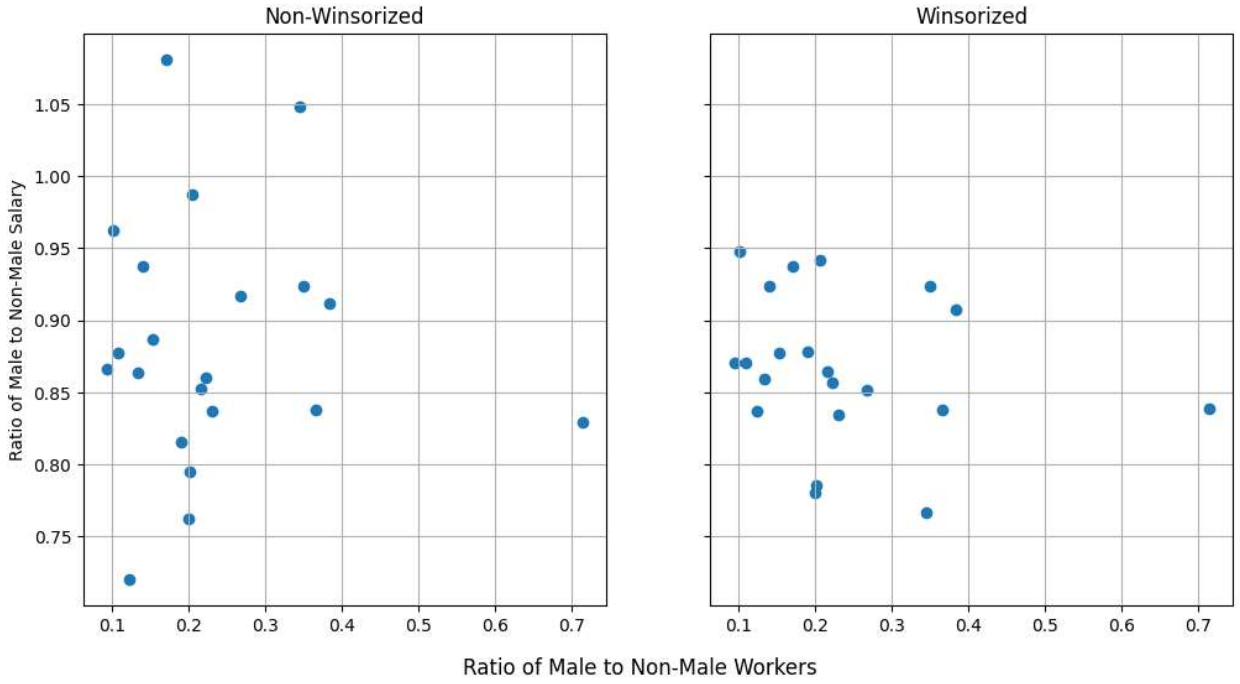
Ratio of Male to Non-Male Workers Compared to Mean Annual Salary



Note: r (18) = .52, $p < .05$ for the non-Winsorized data and *r* (18) = .53, $p < .05$ for the Winsorized data.

Figure 5

Ratio of Male to Non-Male Workers Compared to Ratio of Non-Male to Male Salaries



Note: r (18) = -.003, *p* = .98 for the non-Winsorized data and *r* (18) = -.21, *p* = .35 for the Winsorized data.