

# The Gender Pay Gap in the United States Tech Industry is Largest for Women with 20 to 30 Years of Professional Coding Experience\*

Morgaine Westin

25 April 2021

## Abstract

Despite promises from prominent companies to improve representation and diversity in the workforce, representation of women and racial minorities in the tech industry continues to be a prominent issue. A serious consequence of gender and racial inequalities in the workforce is the wage gap, in which women and women of colour in particular receive lower salaries than their male counterparts. Using data from the 2020 Stack Overflow Developer Survey, multiple linear regression was performed to model differences in salaries among tech professionals based on gender, while taking into consideration the effects of how gender may interact with other factors such as ethnicity and years of experience. Although there were no differences in reported salaries between men and women with less years of professional coding experience, women with 20-30 years of experience reported significantly lower salaries than men with similar levels of experience, highlighting concerns about long-term retention for women in the tech industry and the underrepresentation of women in senior-level positions.

**Keywords:** gender inequality, technology industry, wage gap, race, intersectionality, retention, propensity score matching, multiple linear regression

## 1 Introduction

In 2019, the United States tech industry had 12.1 million employees and nearly 4.6 million job postings, many of which were for new emerging tech areas (CompTIA 2020). As the tech industry continues to grow and reports indicate the median wage for tech professionals is almost double the median national wage in the United States, more people are looking to enter the field. However, despite industry growth and the number of opportunities available, the tech field has been consistently dominated by white or Asian men (Harrison 2019). Gender and racial disparities in science, technology, engineering, and medicine (STEM) fields have been well documented. Notably, women and people of colour, particularly those identifying as African American, Latinx or Hispanic, or Native American, have been disproportionately underrepresented in these fields. (Jackson, Starobin, and Laanan 2013). In response to growing calls for diversity, large tech companies such as Google, Microsoft, and Facebook have launched initiatives and pledges to help support underrepresented groups both within the workplace, and externally through funds, outreach programs, and diverse recruitment practices (Harrison 2019; Banjo and King 2020). Whether or not these initiatives are effective is up for debate, with critics calling some diversity pledges “hollow” and arguing that companies need to make more meaningful efforts before we start seeing significant diversity improvements in the data (Rooney and Khorram 2020; Twine 2018).

One contributor to the diversity problem in tech is retention, or the rate at which underrepresented groups leave the field. Retention and the “leaky pipeline” are frequently discussed in the literature surrounding representation in STEM education, where many women and racial minorities express interest in STEM fields,

---

\*Code and data are available at: <https://github.com/westinmo/2020-dev-survey-analysis>.

but end up switching to different programs early on, with very few individuals in these groups going on to pursue graduate degrees or professional positions (Asai 2020). This extends to the professional world as well, where women and underrepresented racial minorities take up a reasonable portion of entry-level or new tech positions, but are more likely to leave their jobs, and less likely to hold more senior positions in the company (Tapia and Kvasny 2004). Results from a 2017 report which examined why people voluntarily left their jobs in tech found that many women and people of colour were frequently passed over for promotions, and many experienced stereotyping and discrimination in the workplace (Scott, Klein, and Onovakpuri 2017). This suggests company diversity initiatives should not only focus on diverse recruitment and hiring practices, but on creating inclusive and safe workplaces while generating more opportunities for underrepresented groups to grow within the company.

Representation of women in tech in particular has been widely discussed. Among the various STEM fields, the gender gap in computer science is one of the most apparent, where the number of women graduating with a computer science or engineering degree has been decreasing since 1983 (Singh et al. 2007). Even so, women who persist in these degrees often face more challenges when trying to enter the tech industry. A survey conducted by the organization Girls Who Code which aimed to characterize the challenges faced by college-aged women when applying for technical positions found over half of female respondents reported a negative experience, or reported knowing a woman who has had one (Girls Who Code 2019). Many of these women received dismissive, demeaning, and gender-biased remarks in the interview process, with some women experiencing sexual harassment in the form of inappropriate sexual comments and flirting (Girls Who Code 2019). Issues such as these carry on past the early hiring stages and into toxic and sexist workplaces, causing many women to leave their positions due to concerns such as sexual harassment (Scott, Klein, and Onovakpuri 2017).

Much of the research conducted on representation in STEM has focused on the experiences on white, middle class women, and has failed to take an intersectional approach to address how factors such as gender, race, and class interact to put certain groups at a greater disadvantage (Alfrey and Twine 2017). Women of colour face a “double bind” when it comes to representation in STEM fields, and are more likely to fall between the cracks of intervention programs designed to improve diversity (Malcom and others 1976). For instance, programs designed to increase representation for racialized groups are often geared towards men, while programs designed to increase female representation are often biased towards helping women in majority groups (Malcom and others 1976). Finally, individuals who are transgender or non gender conforming, such as non-binary or genderqueer are also thought to be disproportionately underrepresented in the tech industry, however there is significantly less data and research done to examine the experiences of these individuals in tech compared to other underrepresented groups (Dickey 2015).

A major consequence of underrepresentation and gender and racial disparities in the workplace involves wage gaps. In 2019, a report from the tech job platform Hired found that 60% of men were offered higher salaries than their female counterparts for the same position at the same company, and that women were offered 3% less salary on average (Hired 2019). For many women who discovered they were being paid less than their male counterparts in the same position, the difference in salary was at least \$20,000. Moreover, the wage disparities grow when examining both gender and race. While White women and Asian women earn approximately \$0.97 for every dollar earned by White and Asian men, Black women and Latinx women earn \$0.89 and \$0.91 respectively (Hired 2019). This highlights the need to take an intersectional approach when examining wage gaps, as disparities exist even among women.

In this report, I utilized data from the 2020 Stack Overflow Developer Survey to examine representation in the tech industry, and to look for potential wage disparities based on gender and ethnicity. Multiple linear regression modelling was implemented to test the effects of gender and ethnicity on salary for tech professionals in the United States. The first section of the report provides an overview of the Developer Survey and its main results, particularly regarding the demographic characteristics of the respondents<sup>1</sup>. The next sections discuss the process of building and implementing the regression model used to model salary, and its results. Finally, I discuss the implications of my findings in the broader context of representation and bias in the tech industry, and the limitations of my work and how it can be extended upon. The data

---

<sup>1</sup>Survey responses for full-time tech workers in the United States can be explored here: <https://mwtestin.shinyapps.io/Dev-Survey-Shiny-App/>

was prepared and analyzed in R (R Core Team 2020), primarily using the `tidyverse` (Wickham et al. 2019) package, while this report was compiled using R Markdown (Allaire et al. 2021). A full list of packages used in this analysis and report can be found in the Appendix.

## 2 Data

### 2.1 Overview of the Stack Overflow Developer Survey

Stack Overflow is a popular public platform for individuals to ask and answer questions, share knowledge, or collaborate on a wide range of topics related to computer programming (Sewak et al. 2010). The website boasts an average of 100 million user visits per month, and is used by individuals who code from a wide range of backgrounds, including professionals and enthusiasts alike (Stack Overflow 2021). In 2011, Stack Overflow launched their first Annual Developer Survey to better understand its user base. Since then, the annual survey has continued to grow alongside the website’s popularity, with the most recent 2020 survey garnering close to 65,000 responses from users in over 180 countries. While there is some variation in the questions asked from year to year, the Developer Survey covers a wide range of topics related to the experiences of developers, or other people who code. This includes demographic characteristics, questions about job hunting and satisfaction, compensation, programming experience, and the different languages and libraries developers are using. Every year Stack Overflow posts a report overviewing the main findings of the Developer Survey, as well as the full anonymized dataset to download which is available under the Open Database License (ODbL).

I utilized data from the 2020 Developer Survey<sup>2</sup> to address my research questions regarding representation and income inequality in tech due to the popularity and usage of Stack Overflow among tech professionals and wide reach of the survey. Of course, this survey data is not necessarily representative of all tech professionals, as it was a voluntary survey and mainly limited to Stack Overflow users. Respondents were mainly sourced from onsite messaging, blog posts, email lists, Meta posts, banner ads, and social media posts, which might suggest users who were more engaged with Stack Overflow were more likely to notice the survey links and complete the survey. For the 2020 survey, Stack Overflow made additional efforts to diversify their sample by finding ways to advertise the survey outside of their own channels and target coders who may not frequent their websites, as well as promote the survey and provide outreach to underrepresented coders. According to the official report, this resulted in slightly more responses from underrepresented groups compared to previous year, however Stack Overflow acknowledged they still have work to do to increase representation in their sample.

The publicly posted dataset for the 2020 survey contains 64461 observations with 61 variables containing answers to the survey questions. Free response questions and questions with personally identifying information were not included in the dataset. A number of responses that Stack Overflow deemed “unqualified” for analysis were removed from the dataset, namely those where respondents took less than 3 minutes to complete the entire survey, which was estimated to take around 20 minutes to complete. Given the size of the dataset and range of respondents, I narrowed down the responses to focus on salaries for tech professionals living in the United States who are employed full time<sup>3</sup>. Of the 9765 respondents who reside in the United States and are full-time employees, 6368 also reported their income, gender, and ethnicity. Potential response bias should be considered when interpreting the results based on this data, as there may be underlying differences between respondents who reported these characteristics and those who did not. This affects both the extent to which the data is representative of tech professionals in the United States, as well as the generalizability of the model and its results.

---

<sup>2</sup>The overall results of the 2020 survey including more details about the survey’s methodology can be found here: <https://insights.stackoverflow.com/survey/2020>

<sup>3</sup>The survey ran in February of 2020 before the World Health Organization officially declared COVID-19 to be a worldwide pandemic, so factors involving employment status and income may have changed since then

## 2.2 Variables

While the data from the Stack Overflow survey covers a wide range of questions about demographics and work characteristics which all provide valuable information about the tech industry, I focused my analysis on a select few variables of interest which relate to my research question. The variables I will be discussing and focusing on in my model are: salary, gender, ethnicity, years of professional coding experience, developer type and age. When reporting salary, respondents were also asked to indicate whether the salary they input was weekly, monthly, or yearly, as well the currency. These responses were converted to annual salaries in USD by Stack Overflow based on an assumption of 12 months and 50 working weeks. Stack Overflow also trimmed the top 2% highest annual salaries in the dataset and replaced them with a threshold value equivalent to two million USD. However, to help improve model performance by reducing extreme skewness, I removed cases with annual income over \$400,000. In addition to this, I excluded responses where the reported annual salary was less than \$10,000, as many of these responses appeared invalid and were unlikely to reflect true full time salaries. After removing these responses, the overall median salary was \$110,000 USD.

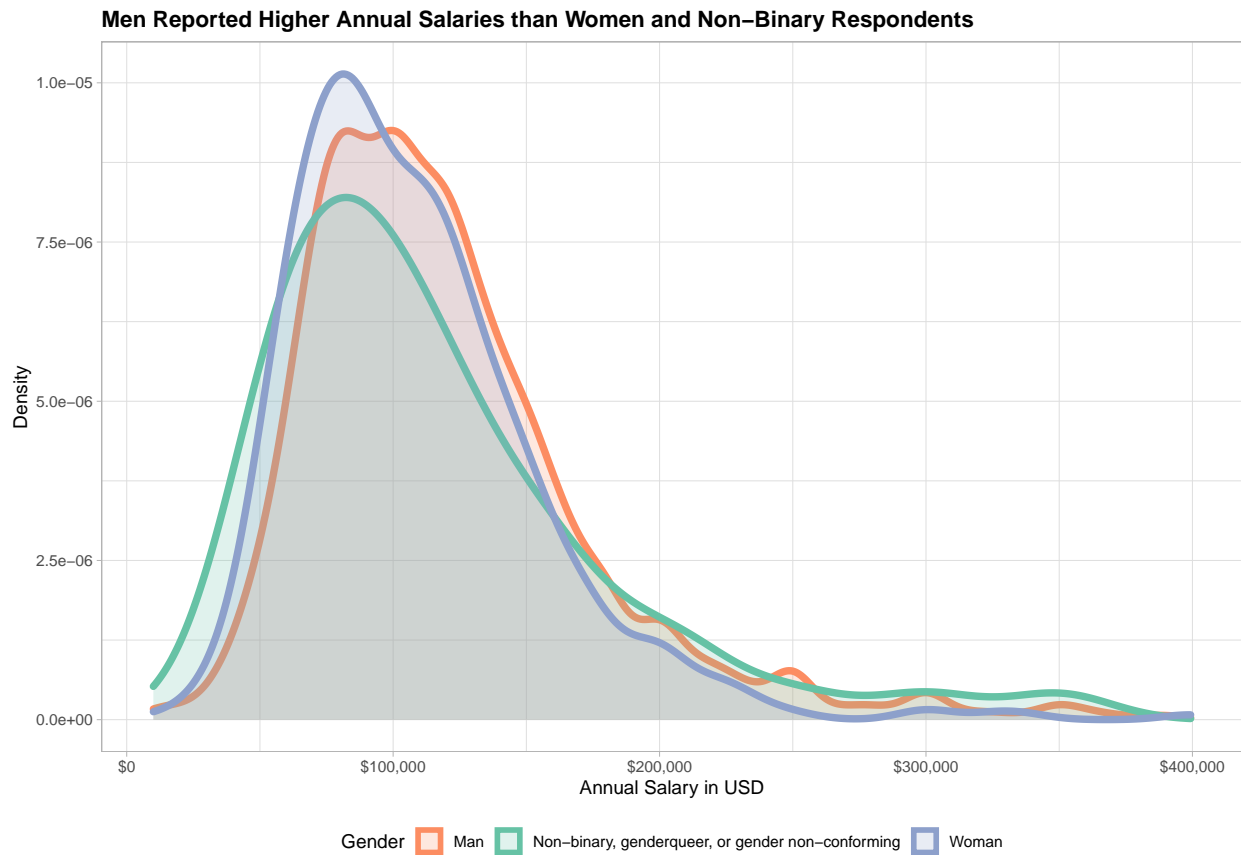


Figure 1: Salary Distributions for Full-Time Developers by Gender

Gender	Respondents	Median	Mean	Standard Deviation
Man	4167	110000	119856.6	54333.32
Non-binary, genderqueer, or gender non-conforming	101	100000	116126.4	64386.71
Woman	496	100000	108263.7	46899.74

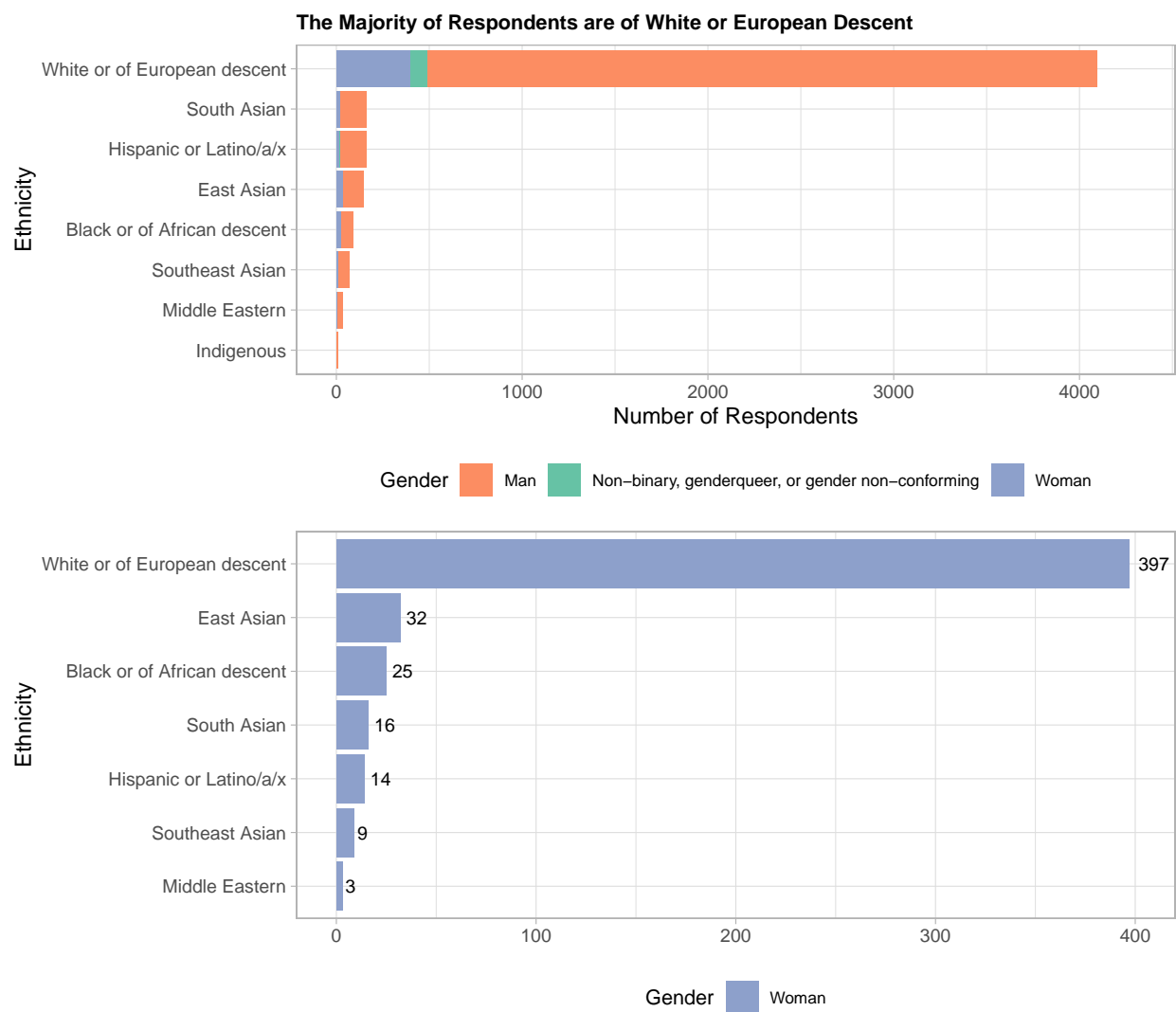


Figure 2: Distribution of Respondents' Ethnicities by Gender

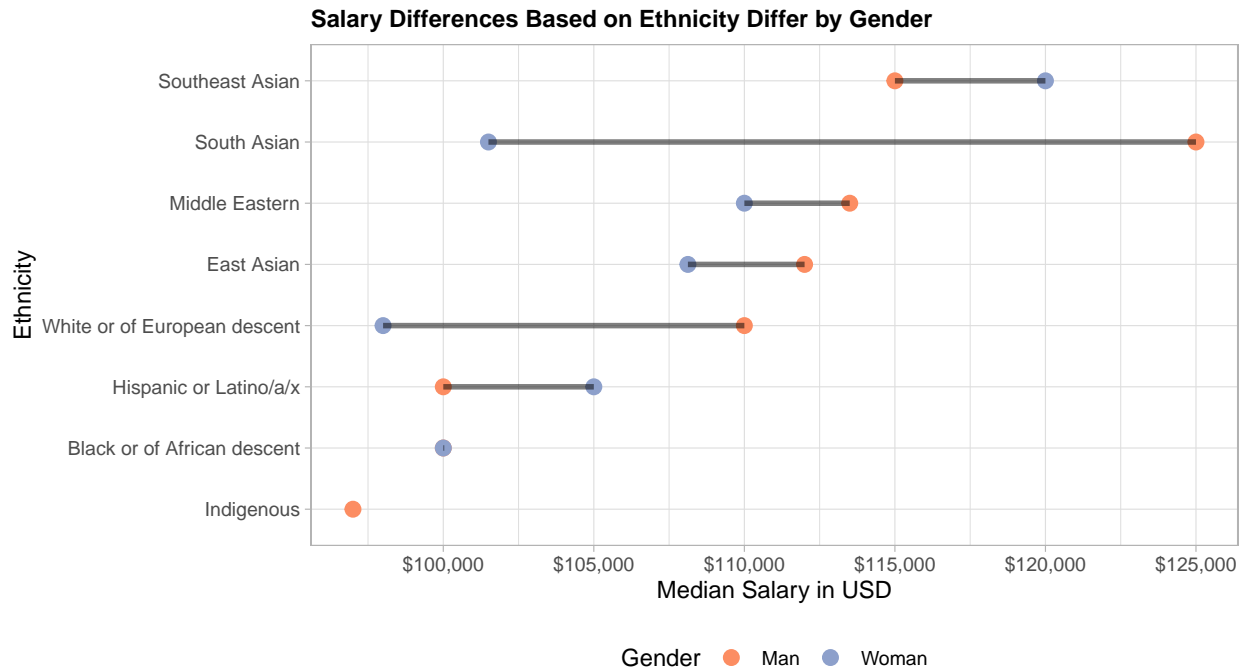
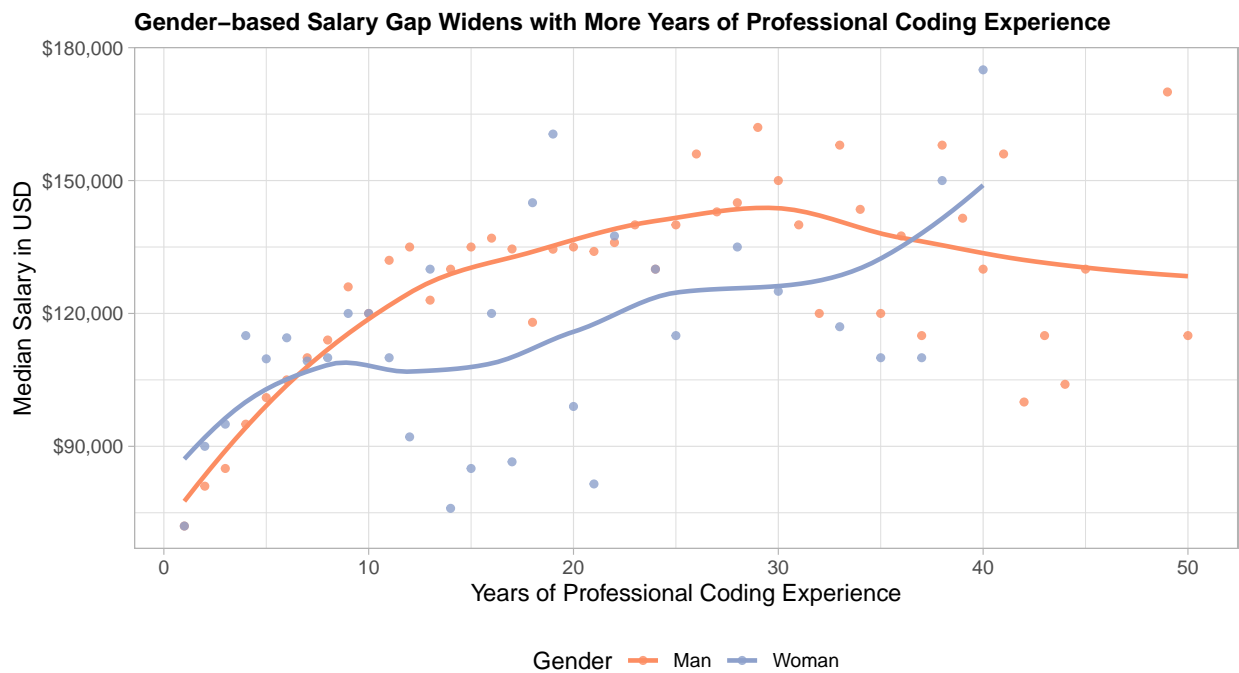


Figure 3: Differences between Median Salaries for Men and Women by Ethnicity



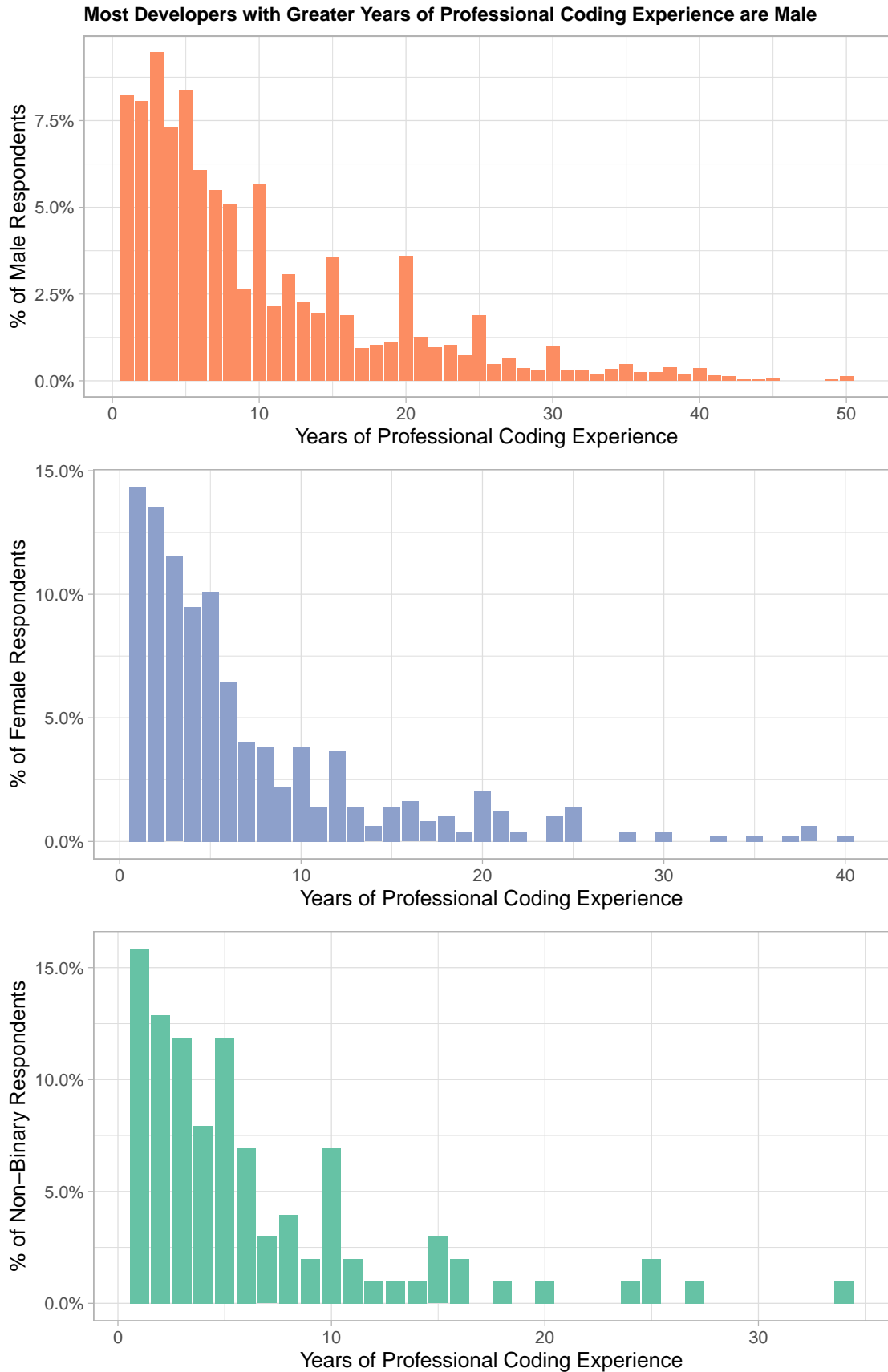
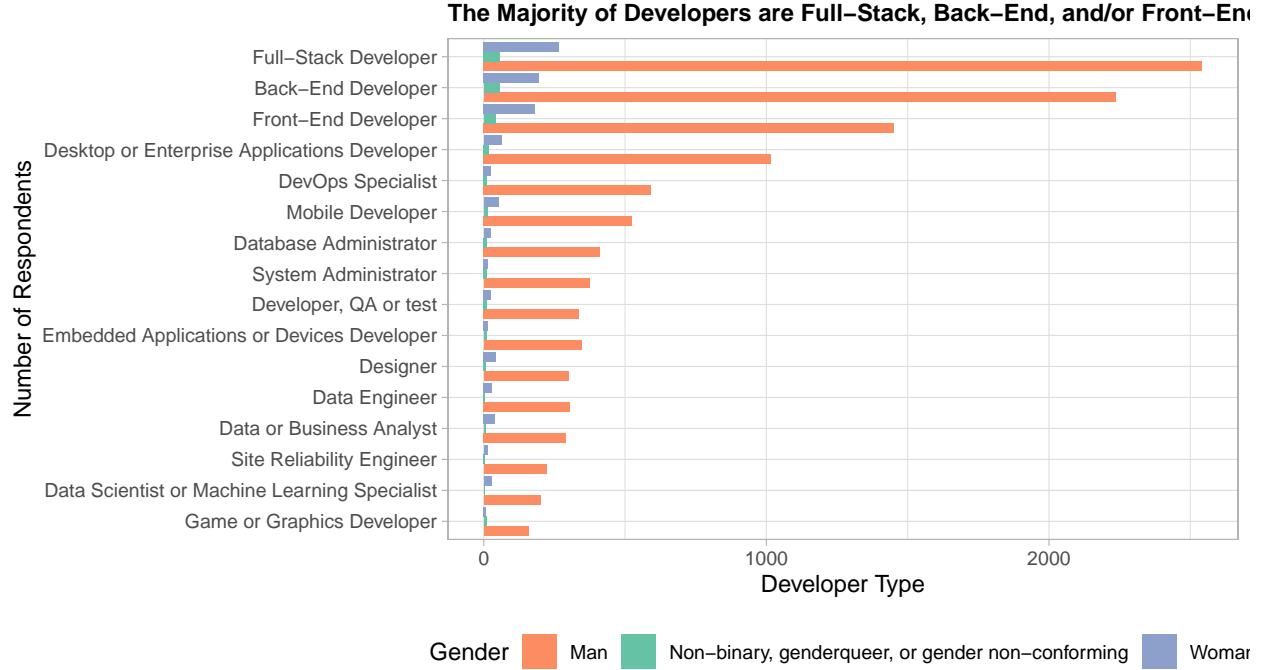


Figure 4: Proportion of Male, Female, and Non-Binary Respondents' Years of Professional Coding Experience



### 3 Model

#### 3.1 Propensity Score Matching

#### 3.2 Multiple Linear Regression Model

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{ethnicity}_i + \beta_3 \text{gender}_i \text{ethnicity}_i + \beta_4 \text{yearscodepro}_i + \beta_5 \text{gender}_i \text{yearscodepro}_i + \beta_6 \text{edlevel}_i + \beta_7 \text{devtype}_i + \beta_8 \text{age}_i + \epsilon_{ij}$$

The output term on the left-hand of the equation represents the response variable, salary. As the salary distribution in the dataset follows a log-normal distribution, a log transformation was applied to the salary variable to normalize the data.  $\beta_0$  is the model's intercept, which represents the mean salary when all predictor values are equal to 0, or categorical variables are at their reference levels. The other  $\beta$  terms in the equation represent coefficients the model will compute for each corresponding predictor variable, which are the partial slopes of each predictor variable, and the response variable, salary. Finally,  $\epsilon_i$  represents the model's residual errors.

- *Gender*
- *Ethnicity*
- *Gender\*Ethnicity*
- *YearsCodePro*
- *Gender\*YearsCodePro*
- *EdLevel*
- *DevType*



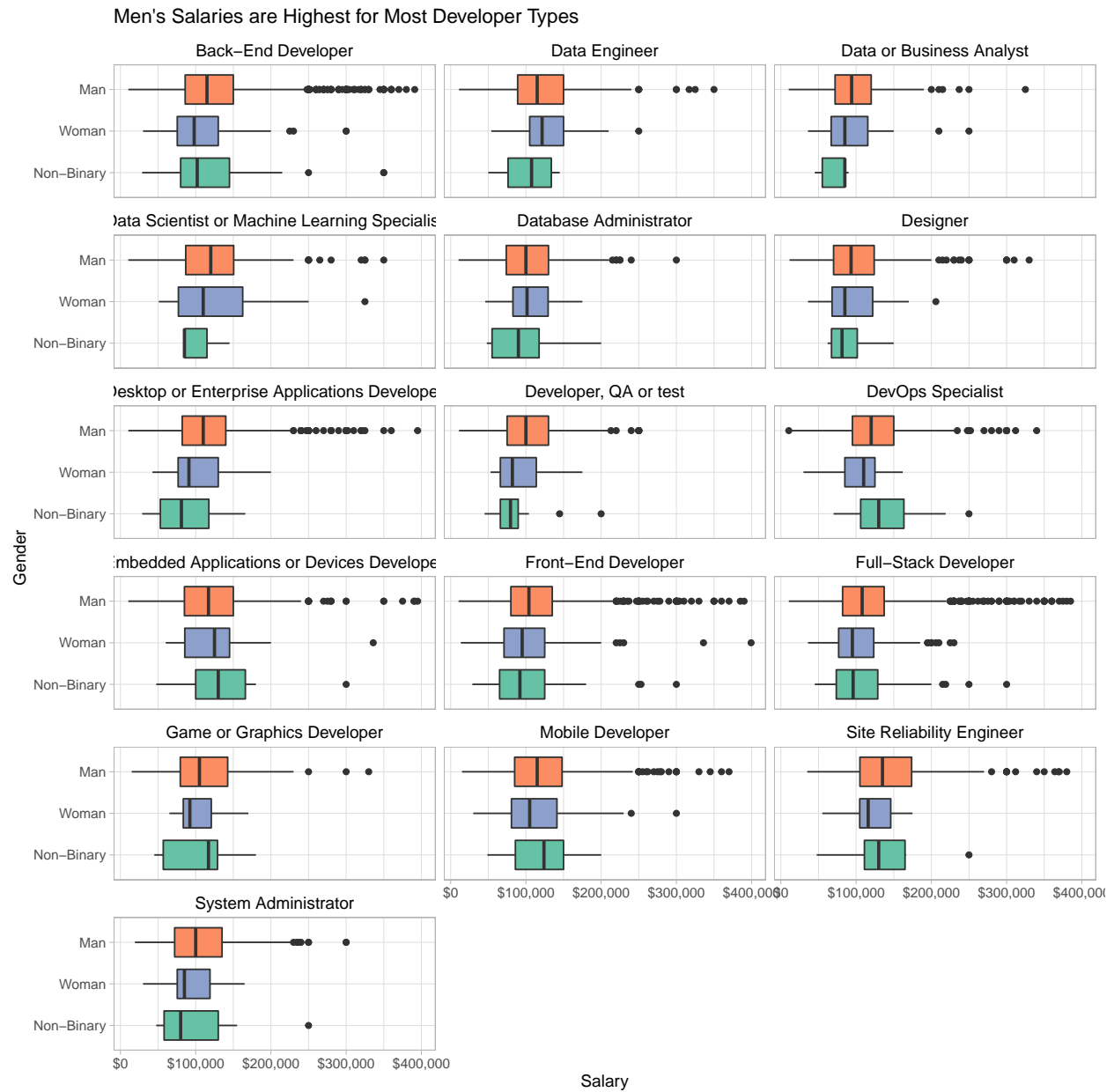


Figure 5: Salary Distributions for Developer Types



Figure 6: Visual Check of Multiple Linear Regression Assumptions

## 4 Results

$$100 * (e^{\beta_j} - 1)$$

i think j is the # of attributes (columns) and i is the number of observations (rows)

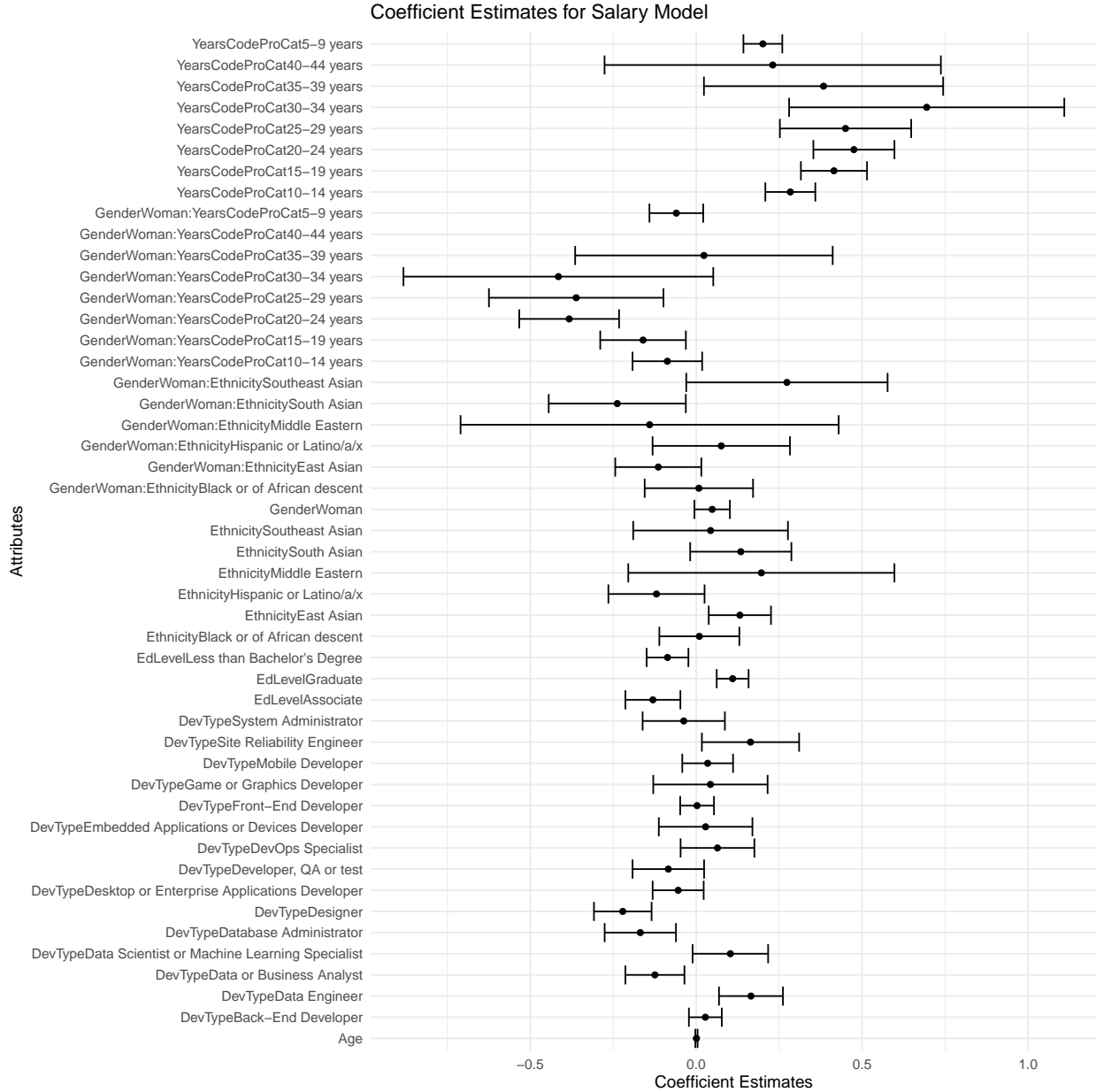


Figure 7: Coefficients for Multiple Linear Regression Model

Characteristic	Beta	95% CI <sup>1</sup>	p-value
<b>(Intercept)</b>	11.323	11.217, 11.428	<b>&lt;0.001</b>
<b>Gender</b>			
Man	—	—	
Woman	0.048	-0.005, 0.101	0.078
<b>Ethnicity</b>			
White or of European descent	—	—	
Black or of African descent	0.009	-0.111, 0.130	0.88
East Asian	0.132	0.038, 0.225	<b>0.006</b>
Hispanic or Latino/a/x	-0.120	-0.265, 0.025	0.11
Middle Eastern	0.196	-0.205, 0.597	0.34
South Asian	0.134	-0.018, 0.287	0.084
Southeast Asian	0.043	-0.190, 0.276	0.72
<b>EdLevel</b>			
Bachelor's	—	—	
Associate	-0.131	-0.213, -0.048	<b>0.002</b>
Graduate	0.110	0.062, 0.158	<b>&lt;0.001</b>
Less than Bachelor's Degree	-0.086	-0.149, -0.024	<b>0.007</b>
<b>Age</b>	0.001	-0.003, 0.004	0.65
<b>DevType</b>			
Full-Stack Developer	—	—	
Back-End Developer	0.028	-0.022, 0.077	0.28
Data Engineer	0.165	0.069, 0.261	<b>&lt;0.001</b>
Data or Business Analyst	-0.124	-0.213, -0.035	<b>0.006</b>
Data Scientist or Machine Learning Specialist	0.103	-0.011, 0.217	0.076
Database Administrator	-0.168	-0.276, -0.061	<b>0.002</b>
Designer	-0.221	-0.308, -0.134	<b>&lt;0.001</b>
Desktop or Enterprise Applications Developer	-0.054	-0.131, 0.022	0.17
Developer, QA or test	-0.084	-0.192, 0.024	0.13
DevOps Specialist	0.064	-0.047, 0.175	0.26
Embedded Applications or Devices Developer	0.028	-0.113, 0.169	0.69
Front-End Developer	0.003	-0.048, 0.053	0.92
Game or Graphics Developer	0.043	-0.129, 0.215	0.63
Mobile Developer	0.035	-0.042, 0.111	0.38

<sup>1</sup>CI = Confidence Interval

R<sup>2</sup> = 0.193; Adjusted R-Squared = 0.172; Sigma = 0.352; Statistic = 8.96; p-value = 0.000; df = 47.0; Log-likelihood = -652; AIC = 1,402; BIC = 1,672; Deviance = 218; Residual df = 1,756; No. Obs. = 1,804

Characteristic	Beta	95% CI <sup>1</sup>	p-value
Site Reliability Engineer	0.164	0.017, 0.310	0.029
System Administrator	-0.038	-0.161, 0.086	0.55
<b>YearsCodeProCat</b>			
Less than 5 years	—	—	
10-14 years	0.284	0.208, 0.359	<b>&lt;0.001</b>
15-19 years	0.415	0.315, 0.515	<b>&lt;0.001</b>
20-24 years	0.475	0.353, 0.597	<b>&lt;0.001</b>
25-29 years	0.450	0.252, 0.648	<b>&lt;0.001</b>
30-34 years	0.695	0.280, 1.109	<b>0.001</b>
35-39 years	0.384	0.023, 0.744	0.037
40-44 years	0.231	-0.276, 0.737	0.37
5-9 years	0.201	0.142, 0.259	<b>&lt;0.001</b>
<b>Gender * Ethnicity</b>			
Woman * Black or of African descent	0.008	-0.155, 0.171	0.92
Woman * East Asian	-0.114	-0.244, 0.015	0.084
Woman * Hispanic or Latino/a/x	0.076	-0.131, 0.283	0.47
Woman * Middle Eastern	-0.140	-0.710, 0.429	0.63
Woman * South Asian	-0.238	-0.444, -0.031	0.024
Woman * Southeast Asian	0.273	-0.030, 0.577	0.077
<b>Gender * YearsCodeProCat</b>			
Woman * 10-14 years	-0.087	-0.192, 0.018	0.10
Woman * 15-19 years	-0.160	-0.289, -0.031	0.015
Woman * 20-24 years	-0.383	-0.533, -0.232	<b>&lt;0.001</b>
Woman * 25-29 years	-0.362	-0.625, -0.099	<b>0.007</b>
Woman * 30-34 years	-0.415	-0.882, 0.051	0.081
Woman * 35-39 years	0.023	-0.365, 0.411	0.91
Woman * 40-44 years			
Woman * 5-9 years	-0.060	-0.141, 0.021	0.15

<sup>1</sup>CI = Confidence Interval

R<sup>2</sup> = 0.193; Adjusted R-Squared = 0.172; Sigma = 0.352; Statistic = 8.96; p-value = 0.000; df = 47.0; Log-likelihood = -652; AIC = 1,402; BIC = 1,672; Deviance = 218; Residual df = 1,756; No. Obs. = 1,804

## 4.1 Overview of Findings

## 4.2 Limitations and Future Directions

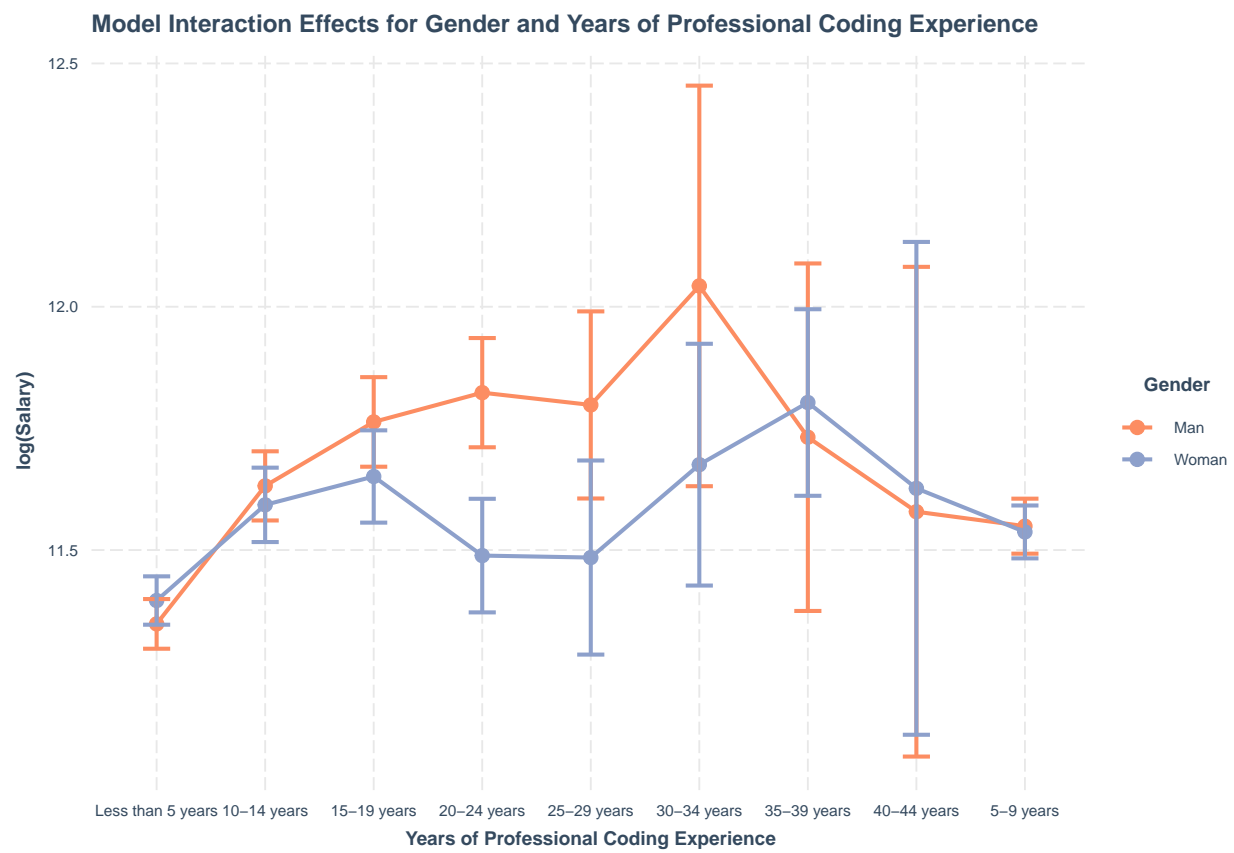


Figure 8: Interaction Plot for Gender and Experience

# Appendix

## Packages

This subsection lists the R packages used in the making of this report.

### General data cleaning and analysis

- `tidyverse` (Wickham et al. 2019)
- `dplyr` (Wickham et al. 2021)
- `janitor` (Firke 2021)
- `purrr` (Henry and Wickham 2020)
- `broom` (Robinson, Hayes, and Couch 2020)
- `scales` (Wickham and Seidel 2020)
- `plyr` (Wickham 2011)
- `forcats` (Wickham 2020)
- `visdat` (Tierney 2017)

### Figures

- `ggplot2` (Wickham 2016)
- `RColorBrewer` (Neuwirth 2014)
- `gridExtra` (Auguie 2017)
- `qqplotr` (Almeida, Loy, and Hofmann 2017)
- `performance` (Lüdecke et al. 2020)
- `interactions` (Long 2019)

### Tables

- `kableExtra` (Zhu 2020)
- `gtsummary` (Sjoberg et al. 2021)
- `flextable` (Gohel 2021)

### RMarkdown Report

- `knitr` (Xie 2021a)
- `bookdown` (Xie 2020)
- `tinytex` (Xie 2021b)

### Shiny App

- `shiny` (Chang et al. 2020)
- `shinythemes` (Chang 2021)
- `plotly` (Sievert 2020)

### Propensity Score Matching and Modeling

- `arm` (Gelman and Su 2020)
- `MLmetrics` (Yan 2016)

## References

- Alfrey, Lauren, and France Winddance Twine. 2017. "Gender-Fluid Geek Girls: Negotiating Inequality Regimes in the Tech Industry." *Gender & Society* 31 (1): 28–50.
- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for R*. <https://github.com/rstudio/rmarkdown>.
- Almeida, Alexandre, Adam Loy, and Heike Hofmann. 2017. *Qqplotr: Quantile-Quantile Plot Extensions for 'Ggplot2'*. <https://github.com/aloy/qqplotr>.
- Asai, David J. 2020. "Race Matters." *Cell* 181 (4): 754–57.
- Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics*. <https://CRAN.R-project.org/package=gridExtra>.
- Banjo, Shelly, and Ian King. 2020. *From Apple to Facebook, Tech's New Diversity Pledges Follow Years of Failure*. Bloomberg. <https://www.bloomberg.com/news/articles/2020-06-23/apple-amazon-facebook-google-microsoft-data-on-black-hiring>.
- Chang, Winston. 2021. *Shinythemes: Themes for Shiny*. <https://CRAN.R-project.org/package=shinythemes>.
- Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2020. *Shiny: Web Application Framework for R*. <https://CRAN.R-project.org/package=shiny>.
- CompTIA. 2020. *Cyberstates 2020*. [https://www.cyberstates.org/pdf/CompTIA\\_Cyberstates\\_2020.pdf](https://www.cyberstates.org/pdf/CompTIA_Cyberstates_2020.pdf).
- Dickey, Megan Rose. 2015. *The Future of Trans\*H4CK*. Tech Crunch. <https://techcrunch.com/2015/11/29/the-future-of-transh4ck/>.
- Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://CRAN.R-project.org/package=janitor>.
- Gelman, Andrew, and Yu-Sung Su. 2020. *Arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. <https://CRAN.R-project.org/package=arm>.
- Girls Who Code. 2019. *Applying for Internships as a Woman in Tech: Findings from a Survey of Gwc-Affiliated Women*. [https://girlswhocode.com/wp-content/uploads/2019/08/GWC\\_Advocacy\\_InternshipApplicationExperiences\\_PDF\\_z6.pdf](https://girlswhocode.com/wp-content/uploads/2019/08/GWC_Advocacy_InternshipApplicationExperiences_PDF_z6.pdf).
- Gohel, David. 2021. *Flextable: Functions for Tabular Reporting*. <https://CRAN.R-project.org/package=flextable>.
- Harrison, Sara. 2019. *Five Years of Tech Diversity Reports—and Little Progress*. Wired. <https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/>.
- Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*. <https://CRAN.R-project.org/package=purrr>.
- Hired. 2019. *The State of Wage Inequality in the Workplace*. <https://hired.com/page/wage-inequality-report>.
- Jackson, Dimitra Lynette, Soko S Starobin, and Frankie Santos Laanan. 2013. "The Shared Experiences: Facilitating Successful Transfer of Women and Underrepresented Minorities in Stem Fields." *New Directions for Higher Education* 2013 (162): 69–76.
- Long, Jacob A. 2019. *Interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions*. <https://cran.r-project.org/package=interactions>.
- Lüdecke, Daniel, Dominique Makowski, Philip Waggoner, and Indrajeet Patil. 2020. "Performance: Assessment of Regression Models Performance." *CRAN*. <https://doi.org/10.5281/zenodo.3952174>.



- Malcom, Shirley Mahaley, and others. 1976. “The Double Bind: The Price of Being a Minority Woman in Science. Report of a Conference of Minority Women Scientists, Arlie House, Warrenton, Virginia.”
- Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes*. <https://CRAN.R-project.org/package=RColorBrewer>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Robinson, David, Alex Hayes, and Simon Couch. 2020. *Broom: Convert Statistical Objects into Tidy Tibbles*. <https://CRAN.R-project.org/package=broom>.
- Rooney, Kate, and Yasmin Khorram. 2020. *Tech Companies Say They Value Diversity, but Reports Show Little Change in Last Six Years*. CNBC. <https://www.cnbc.com/2020/06/12/six-years-into-diversity-reports-big-tech-has-made-little-progress.html>.
- Scott, Allison, Freeda Kapur Klein, and Uriridiakoghene Onovakpuri. 2017. *Teach Leavers Study*. Kapur Center for Social Impact. <https://www.kapurcenter.org/wp-content/uploads/2017/08/TechLeavers2017.pdf>.
- Sewak, M, Divye Raj Khilmani, Andrew Tronson, Kari Okamoto, Shantanu Vikas Kurhekar, Lijing Zhang, Daniel Sanchez, et al. 2010. “Finding a Growth Business Model at Stack Overflow.” *Inc. Stanford Case Publisher*, 1–35.
- Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman; Hall/CRC. <https://plotly-r.com>.
- Singh, Kusum, Katherine R Allen, Rebecca Scheckler, and Lisa Darlington. 2007. “Women in Computer-Related Majors: A Critical Synthesis of Research and Theory from 1994 to 2005.” *Review of Educational Research* 77 (4): 500–533.
- Sjoberg, Daniel D., Michael Curry, Margie Hannum, Joseph Larmarange, Karissa Whiting, and Emily C. Zabor. 2021. *Gtsummary: Presentation-Ready Data Summary and Analytic Result Tables*. <https://CRAN.R-project.org/package=gtsummary>.
- Stack Overflow. 2021. *Stack Overflow Company Page*. <https://stackoverflow.com/company>.
- Tapia, Andrea H, and Lynette Kvasny. 2004. “Recruitment Is Never Enough: Retention of Women and Minorities in the It Workplace.” In *Proceedings of the 2004 Sigmis Conference on Computer Personnel Research: Careers, Culture, and Ethics in a Networked Environment*, 84–91.
- Tierney, Nicholas. 2017. “Visdat: Visualising Whole Data Frames.” *JOSS* 2 (16): 355. <https://doi.org/10.21105/joss.00355>.
- Twine, France Winddance. 2018. “Technology’s Invisible Women: Black Geek Girls in Silicon Valley and the Failure of Diversity Initiatives.” *International Journal of Critical Diversity Studies* 1 (1): 58–79.
- Wickham, Hadley. 2011. “The Split-Apply-Combine Strategy for Data Analysis.” *Journal of Statistical Software* 40 (1): 1–29. <http://www.jstatsoft.org/v40/i01/>.
- . 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- . 2020. *Forcats: Tools for Working with Categorical Variables (Factors)*. <https://CRAN.R-project.org/package=forcats>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolmund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.

- Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. <https://CRAN.R-project.org/package=scales>.
- Xie, Yihui. 2020. *Bookdown: Authoring Books and Technical Documents with R Markdown*. <https://github.com/rstudio/bookdown>.
- . 2021a. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. <https://yihui.org/knitr/>.
- . 2021b. *Tinytex: Helper Functions to Install and Maintain Tex Live, and Compile Latex Documents*. <https://github.com/yihui/tinytex>.
- Yan, Yachen. 2016. *MLmetrics: Machine Learning Evaluation Metrics*. <https://CRAN.R-project.org/package=MLmetrics>.
- Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.