Diversity Concerns and Wage Inequalities in the United States Tech Industry are Persistent[*]

Morgaine Westin

20 April 2021

Abstract

Despite promises from prominent companies to improve 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.

Contents

Keywords: gender inequality, technology industry, wage gap, race, intersectionality, multiple linear regression

[*] Code and data are available at: https://github.com/westinmo/2020-dev-survey-analysis.

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, 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 though 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. 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 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).

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[1] 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

[1] 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

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[2]. 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.

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.

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

```r
survey_clean <- survey_clean[!is.na(survey_clean$YearsCodeProNew), ]
mean(survey_clean$YearsCodeProNew)
```

```
## [1] 9.535842
```

```r
survey_clean %>%
  ggplot() +
  geom_bar(aes(x = YearsCodeProNew, fill = Gender), stat = "count", position = "dodge") +
  theme_light() +
  theme(legend.position = "bottom")
```

[2] 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

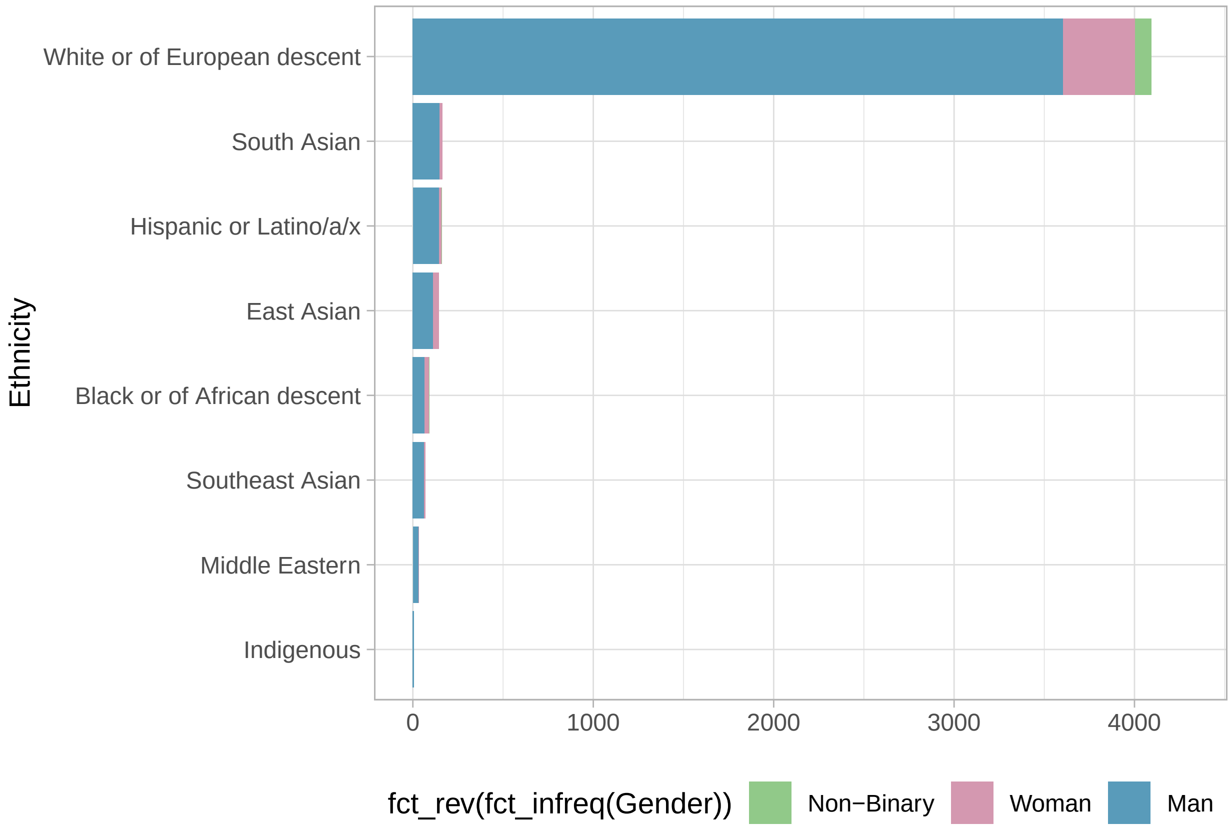

Ethnicity
White or of European descent
South Asian
Hispanic or Latino/a/x
East Asian
Black or of African descent
Southeast Asian
Middle Eastern
Indigenous
0
1000
2000
3000
4000
fct_rev(fct_infreq(Gender))
Non−Binary
Woman
Man

Figure 1: No

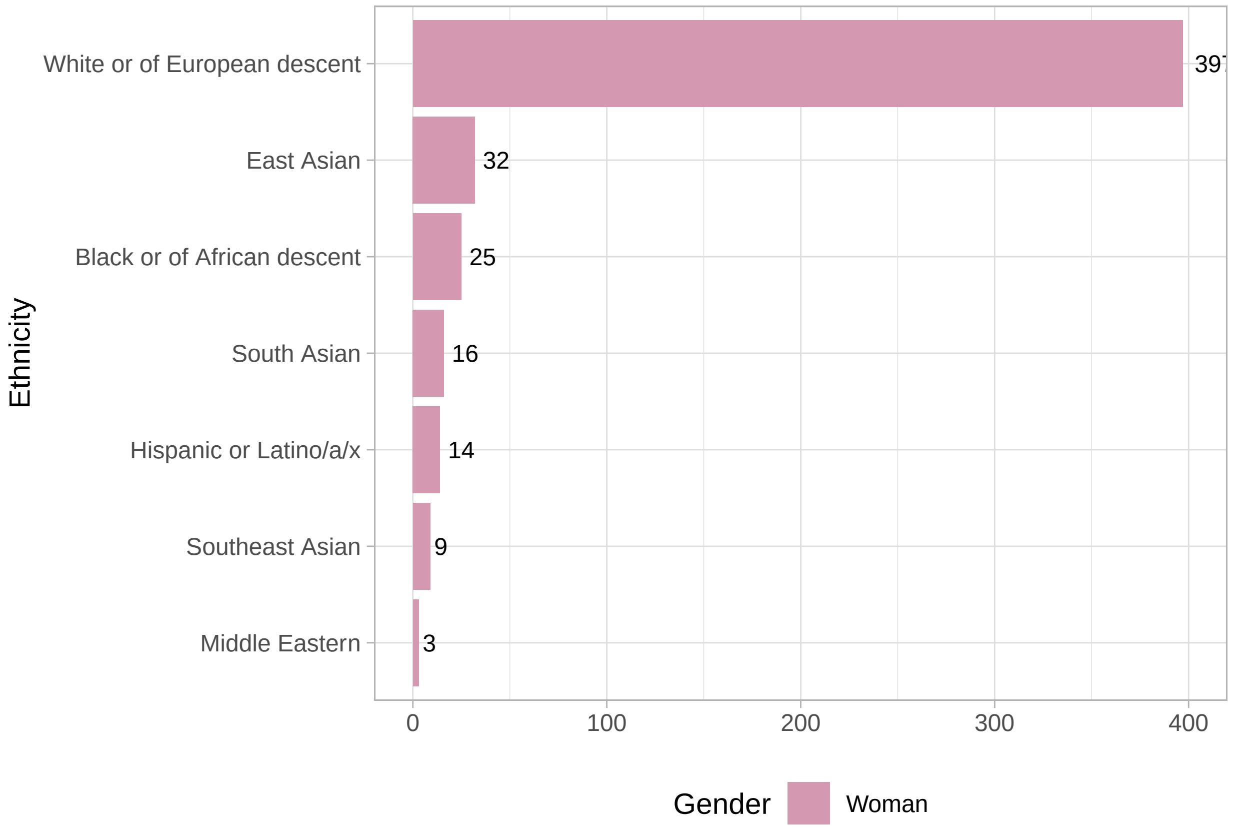


Figure 2: test

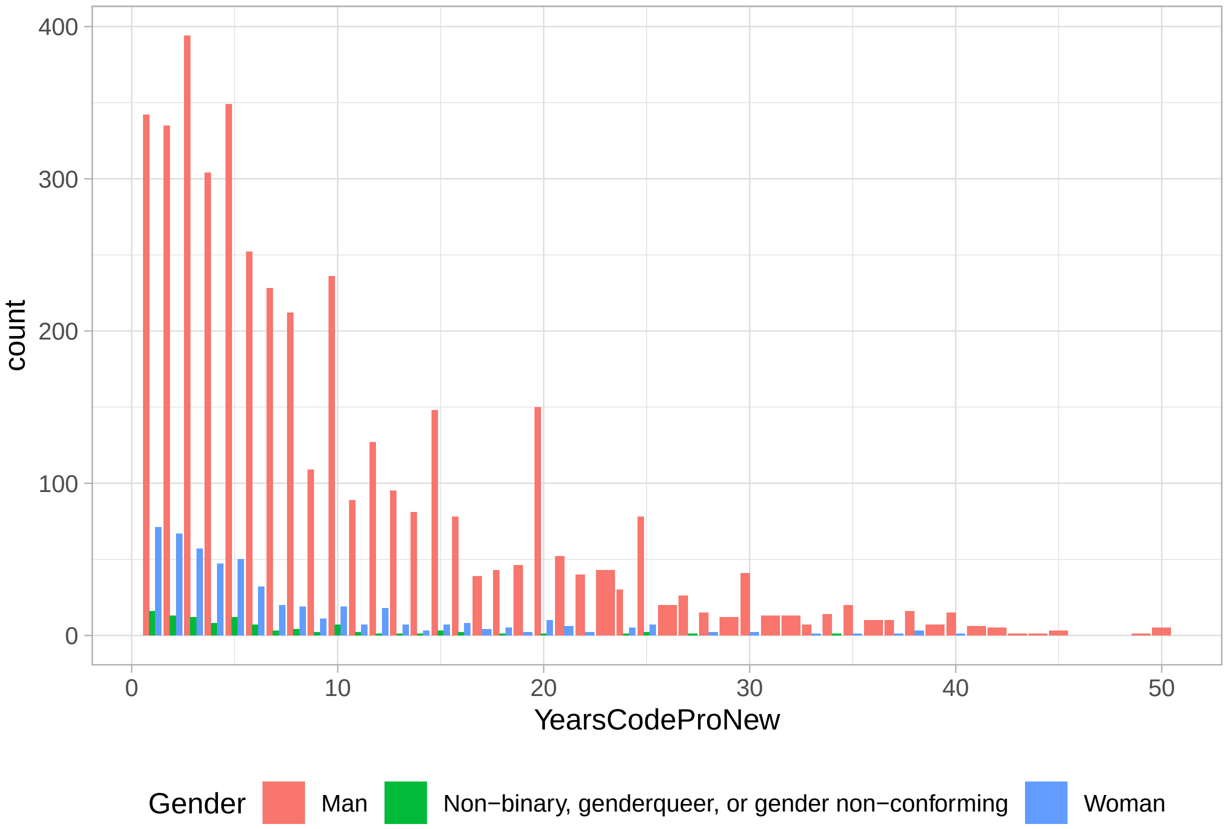


Figure 3: test

```
survey_clean$YearsCodeProCat <- cut(survey_clean$YearsCodeProNew, breaks = c(0,5,10,15,20,25,30,35,40,4
                                      labels = c("Less than 5 years","5-9 years","10-14 years","15-19 y


survey_matched %>%
  ggplot(aes(x = YearsCodeProCat, y = Salary, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge")
```

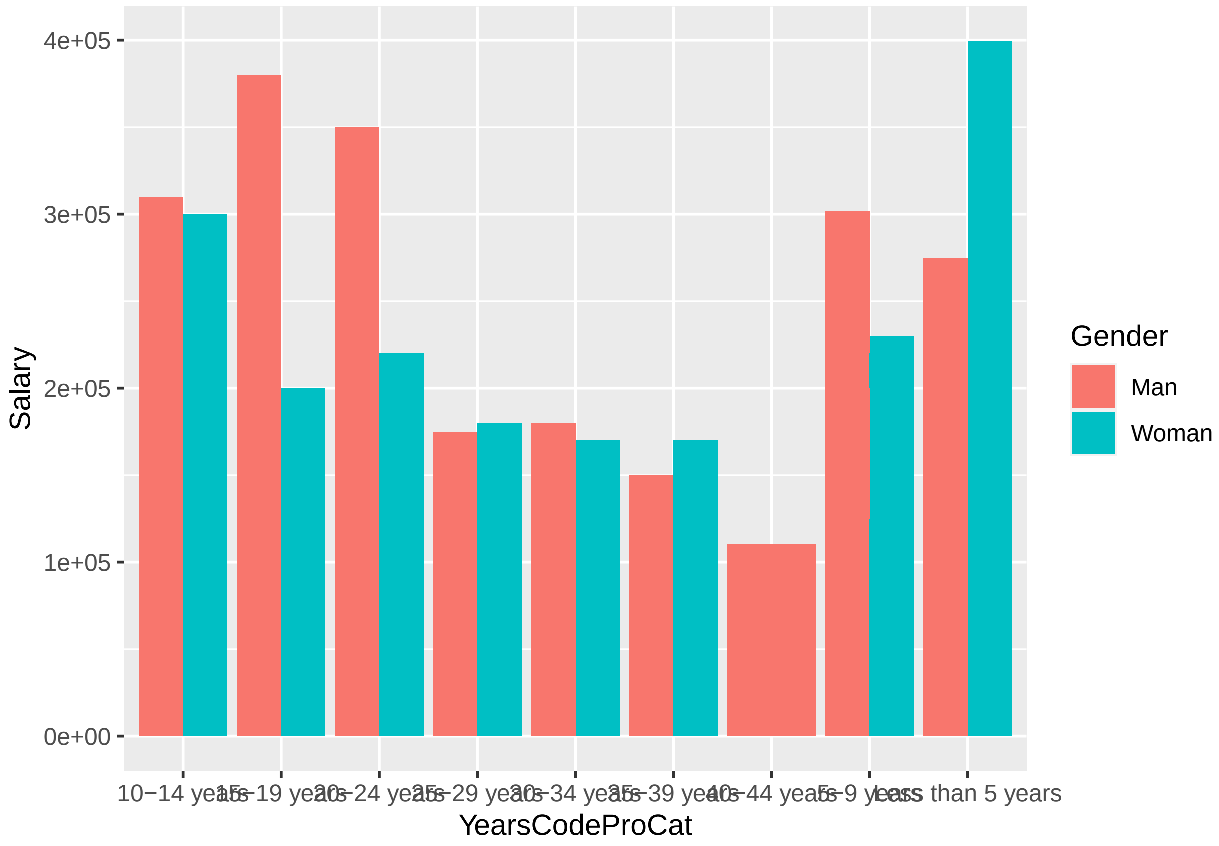


Figure 4: test

```
#survey_clean %>%
  #dplyr::group_by(Gender) %>%
  #dplyr::summarise(mean(YearsCodeProNew))

#survey_clean %>%
  #dplyr::group_by(Gender) %>%
  #dplyr::summarise(mean(Age, na.rm=TRUE))
```

```
survey_unnest %>%
  ggplot(aes(x = fct_rev(fct_infreq(Gender)), y = Salary, fill = Gender)) +
  geom_boxplot(width = .7) +
  facet_wrap(~DevType, ncol = 3) +
  scale_y_continuous(labels = dollar_format()) +
scale_x_discrete(name = "Gender Identity", labels = c("Non-Binary", "Woman", "Man")) +
coord_flip() +
theme_light() +
scale_fill_manual(values = c("#599bba", "#91c989","#d498b0")) +
  theme(legend.position = "none",
        strip.background = element_blank(),
        strip.text = element_text(colour = "black", size = "11"))
```

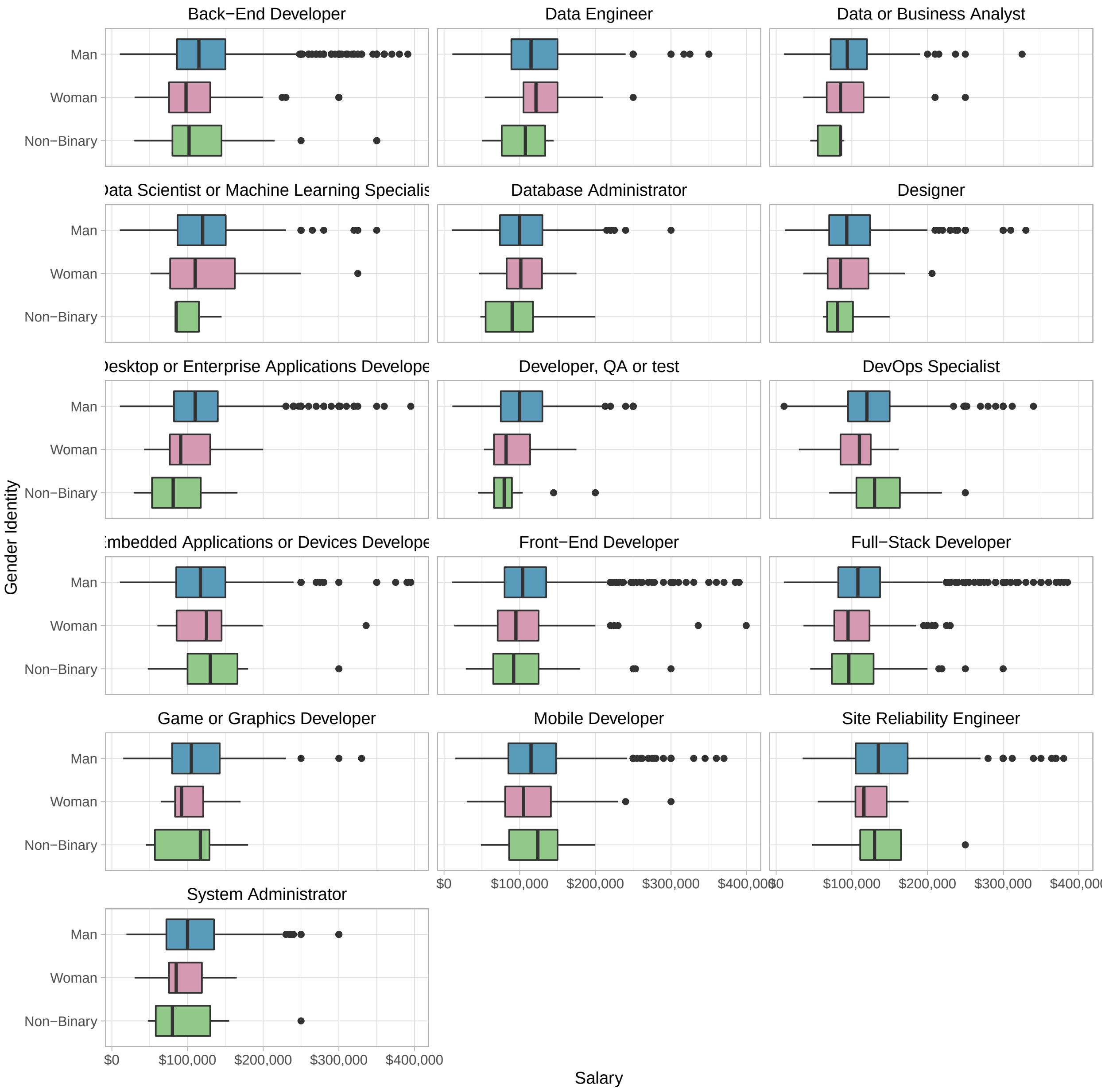


Figure 5: test

```r
survey_unnest %>%
  ggplot(aes(x = fct_rev(fct_infreq(DevType)), fill = fct_rev(fct_infreq(Gender)))) +
  geom_bar(position = "dodge", stat = "count") +
  coord_flip() +
  scale_fill_manual(values = c("#91c989", "#d498b0", "#599bba"), labels = c("Non-Binary", "Woman", "Man
  theme_light() +
  theme(legend.position = "bottom")
```

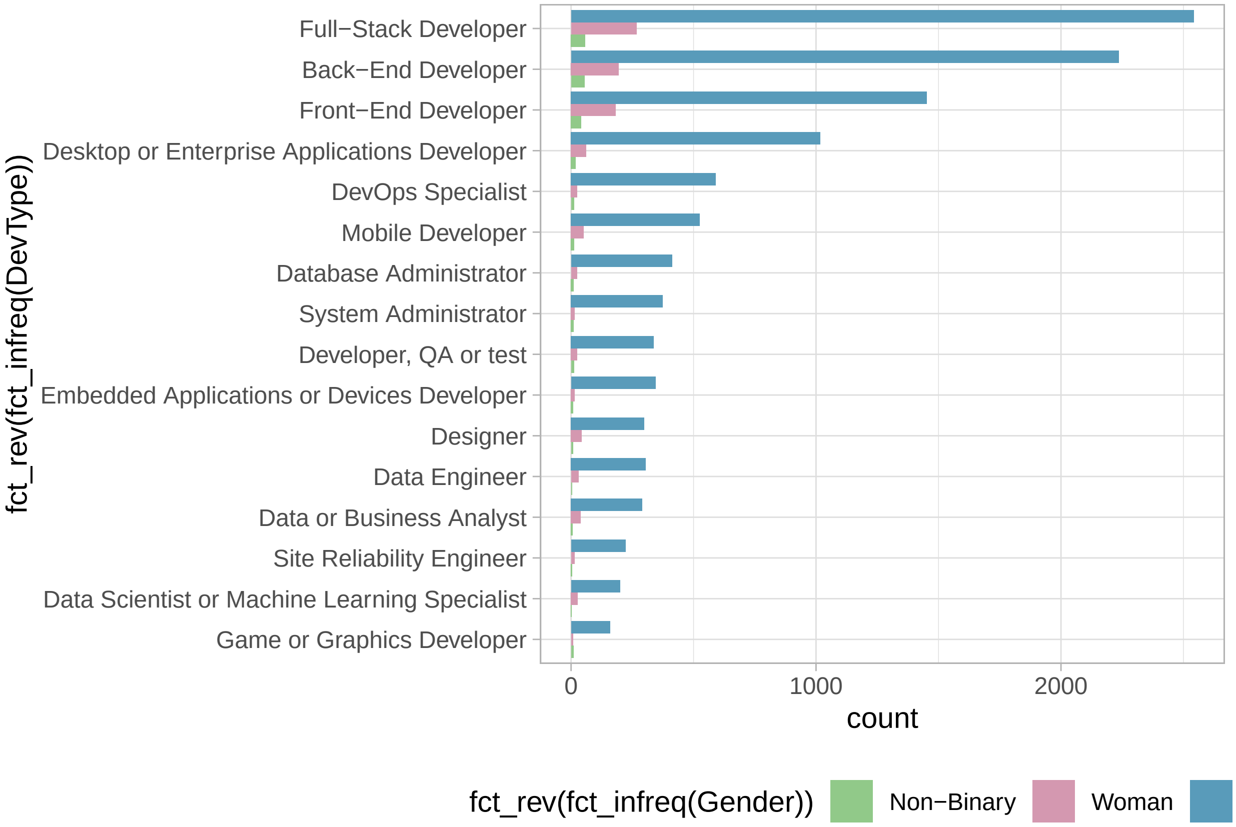


Figure 6: test

3 Model

3.1 Propensity Score Matching

3.2 Model

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{gender} + \beta_2 x_{ethnicity} + \beta_3 x_{test} + \beta_4 x_{state} + \beta_5 x_{education} + \beta_6 x_{hispanic} \qquad (1)$$

```r
check_model(salary_model)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1804 rows containing missing values (geom_text_repel).
```

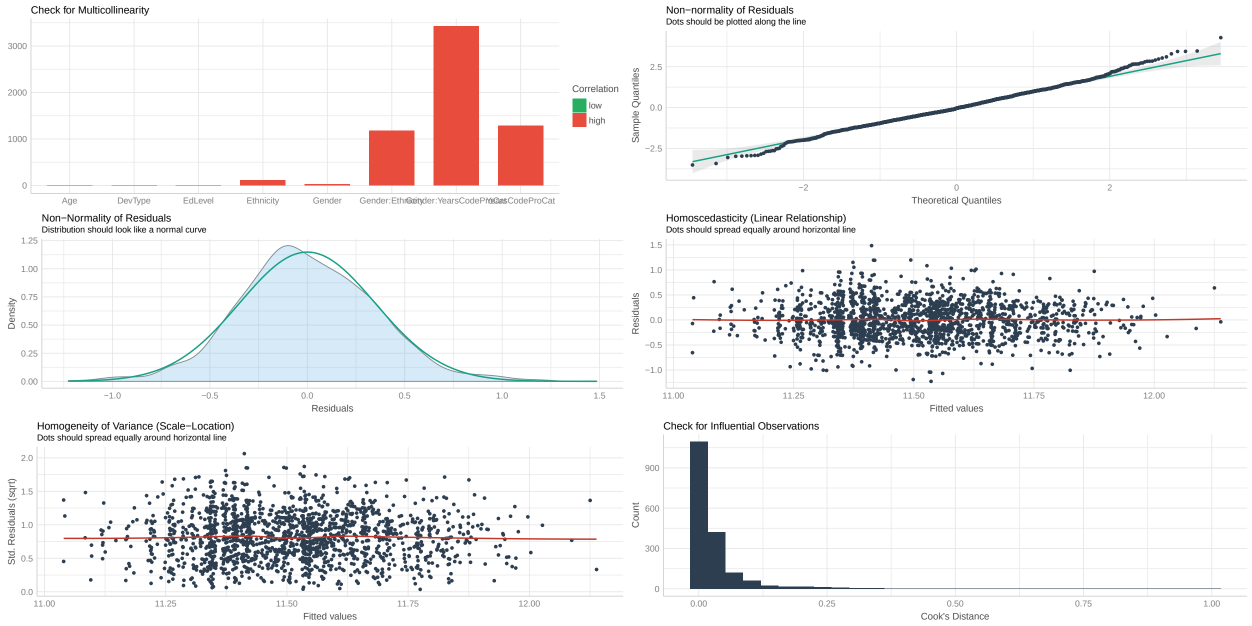


Figure 7: test

4 Results

```
salary_model %>%
  tbl_regression(pvalue_fun = ~style_pvalue(.x, digits = 2), intercept=T) %>%
  bold_p(t = 0.05) %>%
  bold_labels() %>%
  as_flex_table() %>%
  font(fontname = 'Times', part = "all") %>%
  fontsize(size = 10.5, part = "body")
```

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	12	11, 12	**<0.001**
Gender			
Man	—	—	
Woman	-0.03	-0.21, 0.15	0.73

CI = Confidence Interval[1]

Characteristic	Beta	95% CI[1]	p-value
Ethnicity			
Black or of African descent	—	—	
East Asian	0.12	-0.03, 0.27	0.11
Hispanic or Latino/a/x	-0.13	-0.31, 0.06	0.17
Middle Eastern	0.19	-0.23, 0.60	0.38
South Asian	0.12	-0.07, 0.31	0.20
Southeast Asian	0.03	-0.23, 0.29	0.80
White or of European descent	-0.01	-0.13, 0.11	0.88
EdLevel			
Associate	—	—	
Bachelor’s	0.13	0.05, 0.21	**0.002**
Graduate	0.24	0.15, 0.33	**<0.001**
Less than Bachelor’s Degree	0.04	-0.06, 0.14	0.39
Age	0.00	0.00, 0.00	0.65
DevType			
Back-End Developer	—	—	
Data Engineer	0.14	0.04, 0.24	**0.006**
Data or Business Analyst	-0.15	-0.24, -0.06	**0.001**
Data Scientist or Machine Learning Specialist	0.08	-0.04, 0.19	0.20
Database Administrator	-0.20	-0.30, -0.09	**<0.001**
Designer	-0.25	-0.34, -0.16	**<0.001**
Desktop or Enterprise Applications Developer	-0.08	-0.16, 0.00	**0.041**
Developer, QA or test	-0.11	-0.22, 0.00	**0.045**
DevOps Specialist	0.04	-0.08, 0.15	0.52
Embedded Applications or Devices Developer	0.00	-0.14, 0.14	>0.99
Front-End Developer	-0.02	-0.08, 0.03	0.37
Full-Stack Developer	-0.03	-0.08, 0.02	0.28
Game or Graphics Developer	0.02	-0.16, 0.19	0.86
Mobile Developer	0.01	-0.07, 0.09	0.86
Site Reliability Engineer	0.14	-0.01, 0.28	0.071
System Administrator	-0.07	-0.19, 0.06	0.31
YearsCodeProCat			
10-14 years	—	—	
15-19 years	0.13	0.03, 0.24	**0.015**

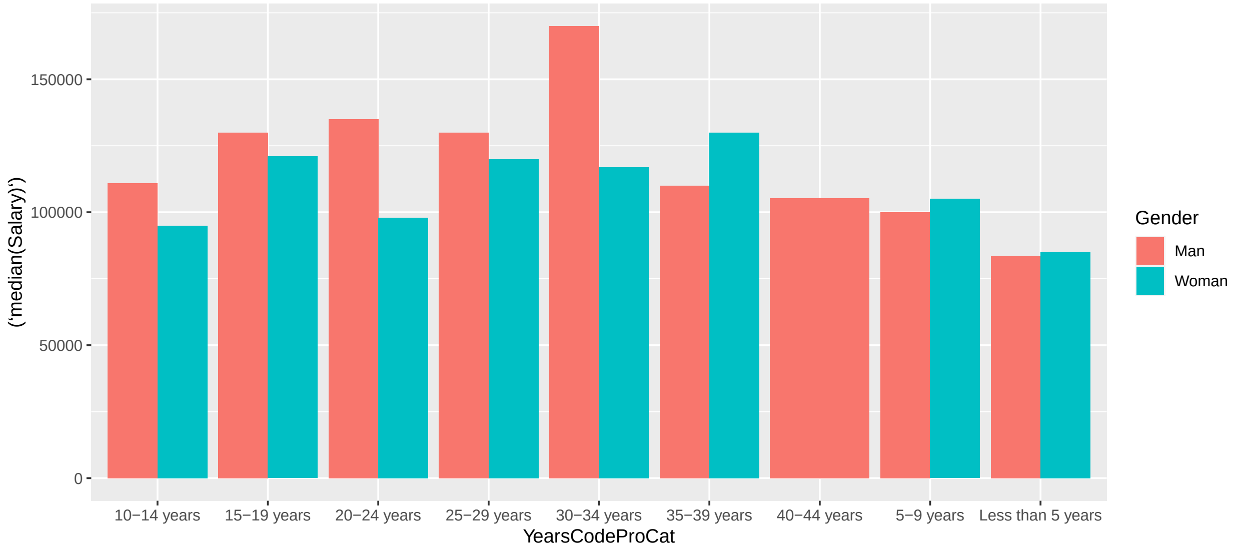
CI = Confidence Interval[1]

Characteristic	Beta	95% CI[1]	p-value
20-24 years	0.19	0.07, 0.32	**0.002**
25-29 years	0.17	-0.03, 0.36	0.10
30-34 years	0.41	0.00, 0.82	0.051
35-39 years	0.10	-0.26, 0.46	0.59
40-44 years	-0.05	-0.56, 0.45	0.84
5-9 years	-0.08	-0.16, 0.00	**0.038**
Less than 5 years	-0.28	-0.36, -0.21	**<0.001**
Gender * Ethnicity			
Woman * East Asian	-0.12	-0.32, 0.08	0.23
Woman * Hispanic or Latino/a/x	0.07	-0.19, 0.33	0.61
Woman * Middle Eastern	-0.15	-0.74, 0.44	0.62
Woman * South Asian	-0.25	-0.50, 0.01	0.061
Woman * Southeast Asian	0.27	-0.07, 0.61	0.13
Woman * White or of European descent	-0.01	-0.17, 0.16	0.92
Gender * YearsCodeProCat			
Woman * 15-19 years	-0.07	-0.22, 0.08	0.34
Woman * 20-24 years	-0.30	-0.46, -0.13	**<0.001**
Woman * 25-29 years	-0.27	-0.55, 0.00	**0.049**
Woman * 30-34 years	-0.33	-0.80, 0.14	0.17
Woman * 35-39 years	0.11	-0.28, 0.51	0.58
Woman * 40-44 years			
Woman * 5-9 years	0.03	-0.08, 0.14	0.63
Woman * Less than 5 years	0.09	-0.02, 0.19	0.10

CI = Confidence Interval[1]

```r
#Graph comparing median salary by experience and gender
survey_matched %>%
  dplyr::group_by(Gender, YearsCodeProCat) %>%
  dplyr::summarise(n(), median(Salary), mean(Salary)) %>%
  ggplot() +
  geom_bar(aes(x = YearsCodeProCat, y = (`median(Salary)`), fill = Gender), stat = "identity", position
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the `.groups` argument.
```



4.1 Overview of Results

4.2 Limitations and Future Directions

Appendix

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.

Asai, David J. 2020. "Race Matters." *Cell* 181 (4): 754–57.

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.

CompTIA. 2020. *Cyberstates 2020*. 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/.

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.

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/.

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.

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."

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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, Freada Kapor Klein, and Uriridiakoghene Onovakpuri. 2017. *Teach Leavers Study*. Kapor Center for Social Impact. https://www.kaporcenter.org/wp-content/uploads/2017/08/TechLeavers2017.pdf.

Sewak, M, Divye Raj Khilnani, 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.

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.

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.

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, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.