Analyzing the Programming Experiences of Women in the Data Scientists and Machine Learning Communities

Kathryn Haglich

12/7/2019

Introduction

For the third year in a row, Kaggle, a popular data science platform, conducted a community wide survey to gain a better understanding of the data science and machine learning community. It is well studied, however, that women have different experiences than their male colleagues primarily due to gender biases. This inequality can even be seen in children as boys are given more programming opportunities and encouragements to persue math and computer science than girls [1]. In a time when data analytics and machine learning are more prominent, it is important to understand the challenges women face and determine ways to support their contributions and participation in the field. The objective of this project is to explore the 2019 Kaggle Data Science and Machine Learning survey and analyze the state of gender inequality in regards to coding experiences with specific attention to years of experience and the number of programming languages regularly used for data analysis.

Data Cleaning and EDA

There were 19,717 respondents from 171 countries who answered the 35 question survey from October 8, 2019 to October 28, 2019. The population consists of individuals connected to Kaggle channels (emails lists, social medial platforms, discussion forums, etc.), and thus does not include individuals who are not part of the Kaggle community.

The survey consisted of 35 multiple choice questions for the participant to answer focusing on demographics, occupational descriptions, and programming opinions and habits. Some questions allowed multiple selection answers with responses to each unique option separated into individual columns. (Question 18, for example, had 12 options to choose from, which is stored as 12 columns in the data set.) Some questions additionally had an option for write-in answers which were stored in a separate data file from the fixed multiple choice answers. Overall, these responses only consisted an extremely small portion of the entire data set and were not considered for the analysis.

The questions that were chosen for this analysis focused on gender (Q2: What is your gender), coding experience (Q15: How long have you been writing code to analyze data (at work or at school)?), and programming language usage (Q18: What programming languages do you use on a regular basis). Overall, the not every question was given to every survey participant, which was determined by previous answers to specific questions. However, the questions of interest were asked to all respondents.

Question 2: Gender

The survey asked participants to identify their preferred gender from the following options: Female, Male, Prefer not to say, and Prefer to self-describe. Those that self described were allowed to write-in the gender they identified as, which was stored as a separate column in the data frame (Q2_Other). The following table shows the frequency of each multiple-choice response.

[1] Cooper, Joel, Kimberlee D Weaver. *Gender and Computers: Understanding the Digital Divide.* 2003.

Gender	Frequency	Proportion
Male	16138	0.818
Female	3212	0.163
Prefer not to say	318	0.016
Prefer to self-describe	49	0.002

This table suggests, with 81% of the respondents identifying as male, that there is significant inequality among all of the genders. Since only 367 out of the 19717 participants responded as neither male nor female, the genders were simplified to the traditional binary factors, female (0) and male (1). Further commentary about the non-traditional gender identifications decoder in the survey can be found in the appendix. The frequency of responses for the two genders is thus as follows:

Gender	Frequency	Proportion
Male	16138	0.834
Female	3212	0.166

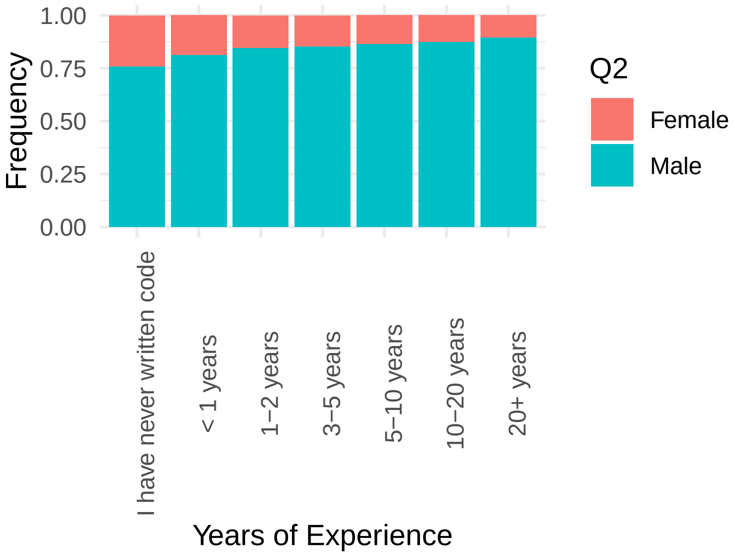
Again, with 83% of the 19350 respondents identifying as male, is a strong indication that there is inequality regarding participation of men and women in the data science and machine learning community.

Question 6: Coding Experience

To measure participants' coding experiences in time, Question 15 asks participants to choose how long they had been using programming languages to analyze data either at school or work. The options and frequency of each category are displayed in the following table:

Current Role	Frequency	Proportion
I have never written code	865	0.055
< 1 years	3828	0.245
1-2 years	4061	0.260
3-5 years	3365	0.215
5-10 years	1887	0.121
10-20 years	1045	0.067
20+ years	576	0.037

This table displays that 72% of survey participants have between 0-5 years of experience analyzing data with programming languages. The graph below visualizes the proportion of the categories answered by each gender.



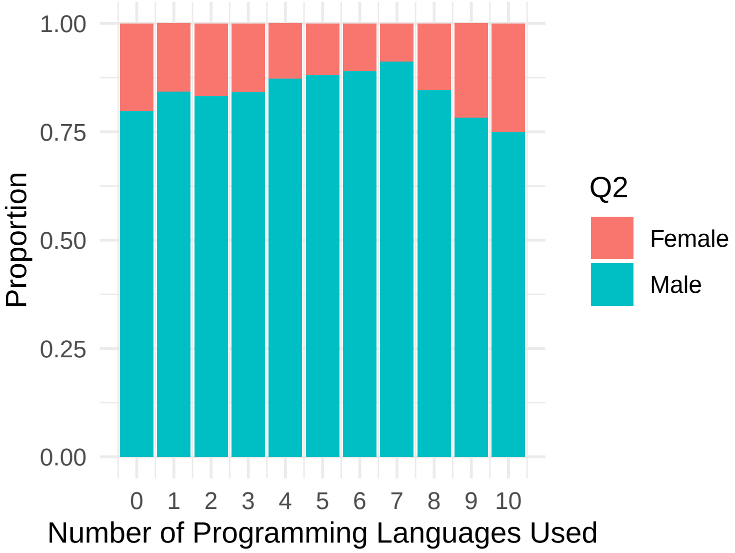
In addition to emphasizing the gender gap first acknowledged in the previous section, this graph indicates that the proportion of women at a given experience level decreases as the experience level advances.

Question 18: Programming Language

The last question considered for this analysis asks which programming languages do people use on a regular basis. As mentioned earlier, this question asked individuals to select as many as they felt applied out of the 12 options available. To gain an overview of quantity of language usage, the total number of languages the person uses, as indicated by which languages were chosen, was calculated. The results of which are seen in the table below.

Total Languages	Frequency	Proportion
0	5396	0.2737
1	3381	0.1715
2	4294	0.2178
3	3527	0.1789
4	1706	0.0865
5	831	0.0421
6	332	0.0168
7	143	0.0073
8	74	0.0038
9	24	0.0012
10	9	0.0005

This table suggests that 27% of individuals do not regularly use a programming language at all with 25% using two. The graph below visualizes the proportion of number of languages indicated by each gender.



This graph indicates out of the 11 categories, the largest proportion of women use no languages at all. Overall, the proportion of women declines (with a slight increase at 2) as the number of programming languages increases until 8 languages where it slightly increases. (However, it should be noted that the frequency of people who use 8, 9, and 10 programming languages regularly is 74, 24, and 9, respectively.)

Models and Results

Experience and Gender

A multinomial regression was implemented to model the relationship between programming experience (y), a 7 level ordered categorical variable, and gender (x), binary variable. This structure model the probability of a specific category occurring based on the response variables (akin to a collection of logistic regression). In ordered multinomial regression, these probabilities are cumulative, and mathematically, it can be represented by the following equation:

$$P(y_i \leq k) = logit^{-1}(\beta_{0k} + \beta_1 x_i)$$

where $k = 1, .., K-1$ and $K =$ the number of categories of the response variables. In this specific model, for example, $P(y = $<1 years) represents the cumulative probability of the outcome being "I have never written code" or "<1 years". For the coefficient, $\beta_1 = 0.36$ with a standard error of 0.04, resulting in the equation,

$$P(y_i \leq k) = logit^{-1}(\beta_{0k} + 0.36x_i)$$

The coefficient represents the log odds, which when exponentiated, leads to an odds value of 1.43. This indicates that, men have higher odds of having more programming experience than women. In particular, for any level of experience, the estimated odds that a man's response is in the higher direction is about 1.43 times the odds for women.

The intercepts vary between the categories as seen in the table below along with the and cumulative probabilities for each gender and category.

Equation	Intercept	Female Probability	Male Probability
P(y <= I have never written code)	-2.548	0.073	0.052
P(y <= < 1 years)	-0.541	0.368	0.289
P(y <= 1-2 years)	0.552	0.635	0.548
P(y <= 3-5 years)	1.555	0.826	0.768
P(y <= 5-10 years)	2.478	0.923	0.893
P(y <= 10-20 years)	3.586	0.973	0.962
P(y <= 20+ years)	-2.548	1	1

To aid in interpretation, the discrete probabilities can be derived from the cumulative probabilities in the following manner, as done in the following table:

$$P(y_i = j) = P(y_i \leq k) - P(y_i \leq k - 1)$$

Probability Equation	Females	Males
P(y = I have never written code)	0.073	0.052
P(y = < 1 years)	0.295	0.237
P(y = 1-2 years)	0.267	0.259
P(y = 3-5 years)	0.191	0.22
P(y = 5-10 years)	0.097	0.125
P(y = 10-20 years)	0.05	0.069
P(y = 20+ years)	0.027	0.038

The overall trends in this table show that the probability that a person will have either less than 1 year or between one and two years of programming experience is highest when compared to the other categories. The probability then decreases as the number of years increases. When considering gender, women are most likely (29.5%) to have less than one year of experience utilizing programming for data analytics. Men are more likely to respond with having one to two years of experience, one category higher than women. It is also important to note that for the lowest categories ("I have never written code" and "less than one year experience"), the probabilities for women are higher than the ones for men. The opposite is true for the remaining categories representing more experiences: men have higher probabilities than women. This indicates that overall, women are more likely to have less programming experience than their male counter parts.

To check the model fit, predicted probabilities based on the model were calculated and used to generate simulated data. The simulation's distribution matches extremely well, as outlined in the appendix, with the observed data, indicating that this is the appropriate model for these variables.

Programming Language and Gender

A Poisson general linear model was implemented to model the total count of programming languages (y) used on a daily basis against gender (x). For this glm, the response variable follows a Poisson distribution, such that

$$y_i \sim Pois(x\beta)$$

and the link function is logistic. The Poisson model for these variables can be represented by the following equations:

$$log(y) = 0.55 + 0.15x$$
$$y = e^{0.55}e^{0.15^x}$$

The intercept and slope have standard errors of 0.014 and 0.015 and 95% confidence intervals of $(0.52, 0.58)$ and $(0.12, 0.18)$, respectively. Since both intervals do not contain zero, these coefficients are considered significant. The mean and standard deviation of the standard residuals are calculated to be $-9.00 * 10^{-18} \approx 0$ and 1, respectively. This, along with a dispersion parameter equal to 1, indicates a very well fitted model. A further analysis of model fitting can be found in the appendix.

When $x = 0$, indicating a woman, $y = 1.73 \approx 1$. Similarly when $x = 1$, indicating a man, $y = 2.014 \approx 2$. Therefore, the following distributions can be concluded:

$$y_f \sim Pois(1) \text{ and } y_m \sim Pois(2)$$

This indicates that the expected number of programming languages used on a regular basis is equal 1 for women and 2 for men. This supports the observations determined by the initial analysis in the previous section: a majority of individuals use between 0-2 languages; the largest proportion of women used either 0 or 2 languages regularly, which averages to 1.

Programming Language, Gender, and Years of Coding Experience

While the models are well fitting individually, it is necessary that the years of coding experience influence the number of languages an individual knows and uses on a regular basis. Thus, it is necessary to consider a multilevel hierarchical Poisson model to account for these differences. Mathematically, the equation would be represented as the following:

$$log(y) = \beta_{0,k} + \beta_1 x$$

where $\beta_{0,k}$ varies by category k. Similarly to the Poisson model, this equation can be rewritten as $y = e^{\beta_{0,k}} e^{\beta_1^x}$ As a result of the nature of the variables (gender as binary and programming experience as 7 level categorical), a Bayesian approach utilizing a Hamiltonian Markov Chain simulation is appropriate to determine the coefficient values. Three chains of 1000 iterations were utilized with a burn-in period of 500 iterations with the default normal prior. Visuals indicating convergence analysis and distributions can be found in the appendix.

The mean value for the slope is 0 with a 90% predictive interval of $(0, 0.1)$. This suggests that there is no relationship between gender and the number of languages regularly used when allowing slope to vary by experience, which contradicts the previous model. Therefore, with this model, only the years of experience has an effect on the number of languages. The mean values of the intercepts are displayed in the table:

Years of Experience	Mean Intercept Value
I have never written code	-0.4
< 1 years	-7.2
1-2 years	1.1
3-5 years	1.2
5-10 years	1.3
10-20 years	1.3
20+ years	1.4

From here, the outcome follows a Poisson distribution, the expected number of languages can be calculated given the experience category, as shown in the table below.

Years of Experience	Expected Number of Languages
I have never written code	0
< 1 years	0
1-2 years	3
3-5 years	3
5-10 years	3
10-20 years	3
20+ years	4

While this model does not give additional information about gender, there is a positive trend between the years of coding experience and the expected number of regularly used programming languages that follows intuition. (Those reporting to have never written code expected to use no programming languages on a regular basis, and those with over 10 years of experience are expected us the most languages.)

Conclusion

Overall, the analysis indicates gender inequality with regards to years of coding experience and the number of programming languages used regularly. The multinomial model shows that shows that women have a higher probability of having less experience utilizing programming for data analysis than men. Why this is the case may be a result of differences in occupation or programming education opportunities between the genders. The Poisson model shows that the expected number of regularly used programming languages for men is higher than the expected number for women. This may also be a reflection of the occupations and education differences: a man with more experience coding may have more opportunities to learn additional languages and be able to be hired in positions requiring multiple languages to be used regularly. This relationship between experience and language usage is most clearly seen in the Bayesian multilevel model, where gender does not influence the model at all. Instead, the experience categories determine the expected number of programming languages; the resulting relationship indicates that those with more programming experience (and thus more opportunities to learn additional languages) have higher expected number of programming languages used on a regular basis.

Future Work

The first significant improvement towards this analysis would be to fine tune the Bayesian multilevel model. Due to time constraints, three chains with 1000 iterations each and a burn-in period of 500 iterations was implemented to keep the algorithm run time at under an hour. This configuration allowed the algorithm to barely reach convergence. (All of the rhat values equal 1, but the lag plots indicate that not all parameters fully converged.) To correct for this, at minimum 4 chains at 2000 iterations each should be implemented to ensure that convergence is obtained providing a more solid foundation for analysis.

Additionally, a model other than a multilevel one might be considered to explore the relationship between the number of languages, programming experience, and gender. A Poisson or negative binomial glm (accounting for interaction) may provide a different perspective between the three variables. These models could also be expanded to account for other factors, such as education level, job title, and size of the company - all of which are included as questions in the survey.

Lastly, this project gives rise to more questions aimed at exploring who are represented in the responses. For example, 27% of the respondents did not use any programming language on a regular people. What occupations or roles are these individuals in? Who are the people with the most experiencing use programming languages for data analysis? Who are the people with the least? What is the distribution of women in these occupations? Which companies are they working for and are there gender based disparities regarding annual compensation? Do women prefer different programming languages than men? Since these variables are all categorical, various multinomial regressions would need to be implemented to explore these inquiries fully.

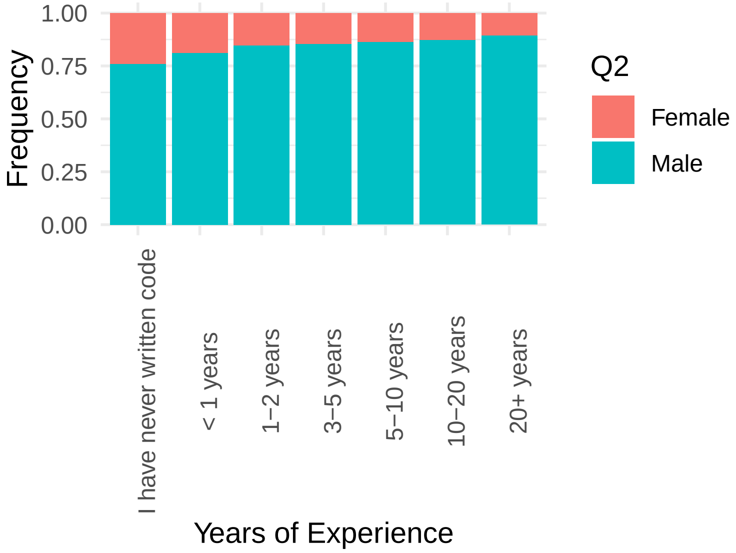
Appendix

Commentary about Gender Representation

As mentioned earlier, the question regarding gender did provide a write-in option, allowing individuals who do not conform with the traditional gender binary to specify their identification. There were a proportionally small amount of individuals who chose to do so (indicating a notable lack of LGBTQ+ representation in the community) and were not included in the analysis of this project. When reviewing the write-in responses, it was noted that there were a significant amount of inappropriate and tongue-in-cheek responses among legitimate gender identities. (Among ones such as "Mister" and "I am a funky potatoes", a prominently offensive was "attack helicopter", and variations thereof, which can be traced to an internet parody.) This tells a story in of itself that inequality does not only happen to women in the field but to those outside of the gender binary spectrum as well. More effort needs to be made to study and understand the challenges LGBTQ+ individuals face in the data science community and to strive for equality.

Model Check for the Multinomial Model

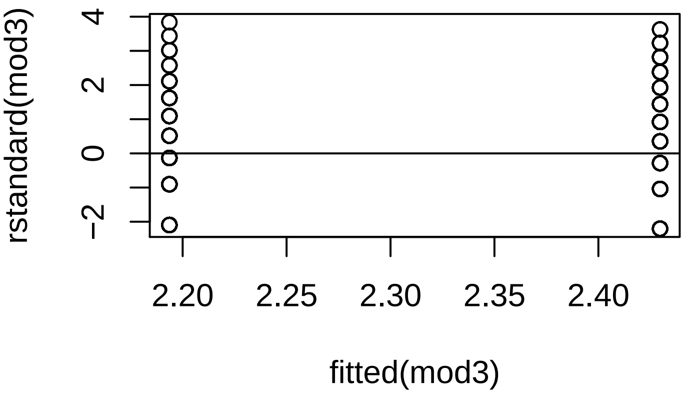
To check the fit of the multinomial regression, a new data set was generated using probabilities predicted from the model. A proportional bar graph similar to the one to visualize the observed relation ship between gender and programming experience was created as shown below.



When compared to to the original graph, the distributions of proportions are nearly identical, thus indicating a good model fit.

Model Check for the Poisson Model

To check the fit of the Poisson model, the following residual plot was created.The pattern observed indicates that there is no apparent overdispertion occurring which indicates that the model is appropriately chosen.



Model Check for the Bayesian Hierarchical Model

Simulation Plots

The plots below are the trace plots, lag plots, distribution curves, and histogram frequencies for the Bayesian multilevel model. The mcmc trace plot indicates that the three chains overall mix well with one or two dramatic jumps in consitent places through out the graphs. The more concerning graphs are the lag plots: only the plots for the slope parameter indicate a quick convergence with the othersrequiring more time. This suggests that not enough iterations were implemented for full convergence to be reached. Despite this overall, the distribution density plots and the histograms provide clear visualizations as to the most likely values of the parameters.