Those who are most likely to work as data scientists

Lu Xu and Zehui(Barry) Zhang

Author contributions

Both of us contribute the design of the project, including direction design, question design, and also the advanced analysis design. To be more specific, Barry did most code works and some analysis work. Lu contributed to the doc more. We complemented each other for the ideas and analysis.

Index

1. Introduction…………………………………………………………………… 1
2. Basic Analysis………………………………………………………………… 2
   2.1 General Glance…………………………………………….…..……… 3
   2.2 Within the Gender…….…………………………………………………..6
   2.3 Working Experiences…………………………………………...…....8
   2.4 Languages…….…………………………………..……………….…...10
3. Advanced Analysis ……………………………………………………………11
4. Conclusion and Discussion …………………………………………………....12
5. Appendix …………………………………………………………………….....13

1.1 Introduction

The designers of this survey are trying to have a comprehensive view of the different country's marching learning and data science. This survey took around 3 years which is from 2017 to 2019 and actually lasts from three and half weeks in October. The final valuable responses are 20,035. Also, this survey is taken without anonymization which is really different from other surveys. Also, this survey provided multiple choices and responses and they could choose all the options that fit on them. Those questions could be what is their highest level of formal education, how many years have they been writing code. If they are not answering the question, it will be reported as an NA. Based on the dataset we could know who is actually working with data, who are moving to other industries and what they are doing in other industries, and what are the ways for the new data scientists to get into the field.

Going along with the direction of gender, this project will start from general to specific, and then make the estimate. With the help of R, we will discover the feathers of gender by associating age, country, level of formal education, roles and the coding experiences. As students and new to data works, we will discover if gender has some relations to participants like us. And mostly, so as to make the work and relationships more apparent and straightforward, we will utilize graphical analysis more than numerical. But we will cover all of both powerful methods. For our advance analysis we will use ANOVA test to check whether the means are the same for the number of people who use different programming languages on a regular basis.

Data and Data Processing

There are 20,035 observations and 355 columns in the original data set and each of the columns represents an answer of a participant for one of thirty-nine questions. Those questions include ages, some basics, genders, and selective questions. Blanks are just the questions that they are not answering.

So as to make the analysis work smooth and easier, we will fill all blanks with NA and mostly focus on the first 20 columns. By using R, remove the first row which represents the details of questions, not the useful data. So the total observations are 20,035. And create two subsets based on Man and Woman so as to compare them easier when we go to specifics.

2. Basic Analysis

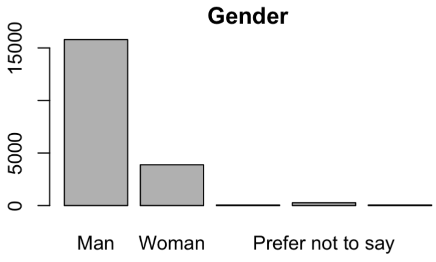
2.1 General Glance

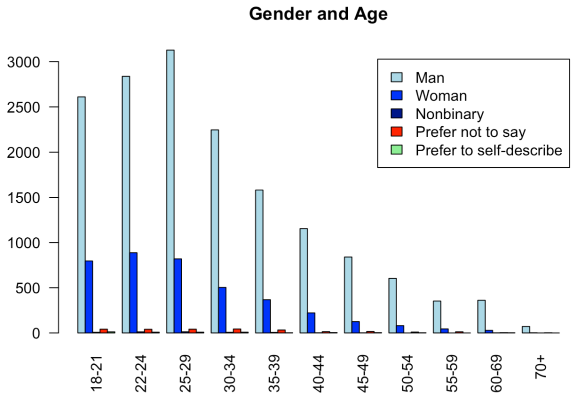
Generally count the number of participants in different genders, and associate it with age, country, and level of education. And show and analyze graphically.

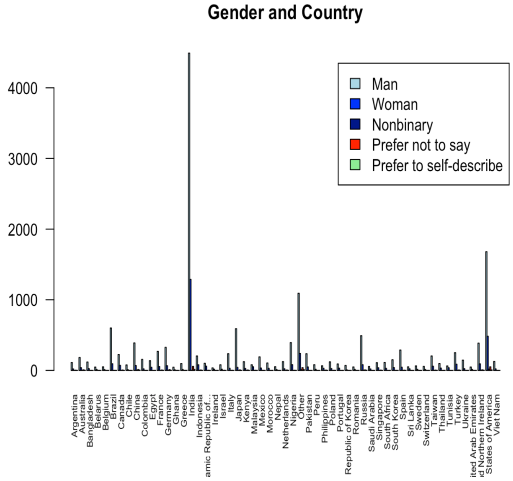
Method

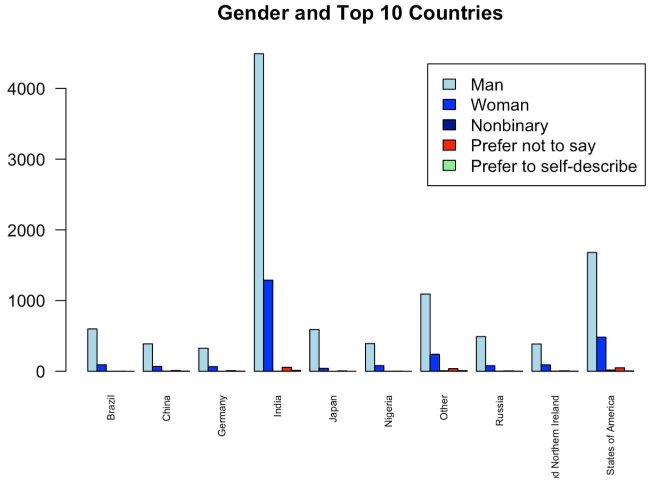
In R, we make a general division between the age, country and level of formal education for 20,035 responses and see their gender. Then we will compare the gender in different categories to make a comparison between male and woman. Graphing shows it and analyzes them.

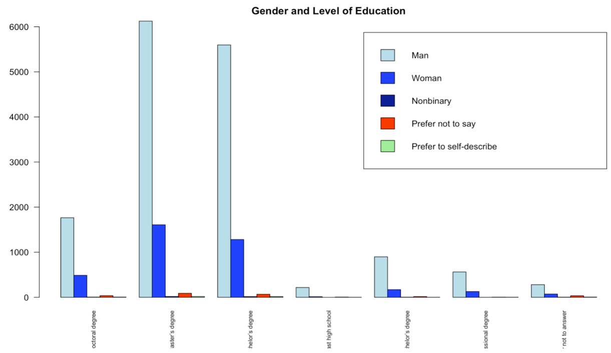
Analysis











From 20,035 responses, we could see that 15789 responses are answered male, and 3878 responses are answered woman, which left 369 response answered nonbinary or prefer to self-dsecribe. Out of 20,035 responses only 369 are not being selected male or female, it is only around 1.8 percent of people so this would not affect the data too much. The number of men in this survey are around 4 times larger than the number of women. By comparing the gender and age, we see that the most common age for women is 22 to 24. The men's slightly older, 25 to 29 years old. They have similar shapes and are skewed to the right which means that there are less old people participating in the survey from the second graph on the top. The main population of the responses currently reside in India, States of America, Other, Brazil, Japan, Russia, Nigeria, Northern Ireland, and China. The most people who are taking this survey currently reside in India and States of America. From the graph the male participants always have the highest number of participants, especially Japan. For the gender and level of education, the bachelor degrees are the most compared to the others. The male still have higher participants than women in this graph too.

Conclusion

From the comparison of different divisions to gender, it shows that the woman has more participants at a younger age than male. Most participants are coming from India, States of American,Brazil, Japan, Russia, Nigeria, Northern Ireland, and China. There are also portions that are from other counties. We could see that Japan actually has more participants in male than women participants. The education levels are really similar. Overally, the male have more than females in all categories.

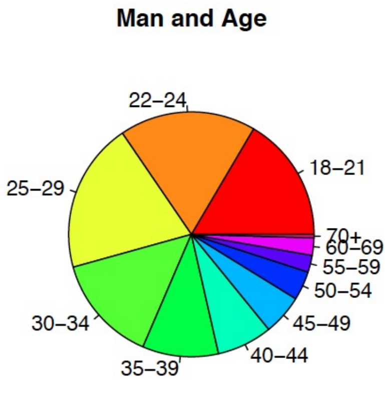
2.2 Within the Gender

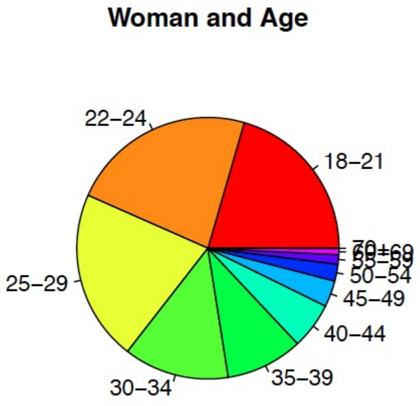
After the analysis in the prior part, we found that Man and Woman almost dominate the proportion in all comparisons. Thus, we narrow our direction to Man and Woman only.

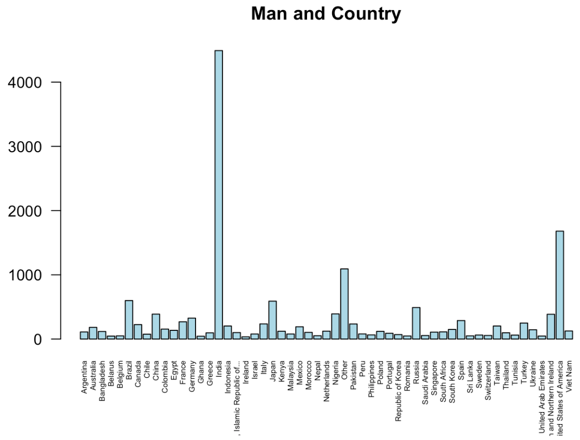
Method

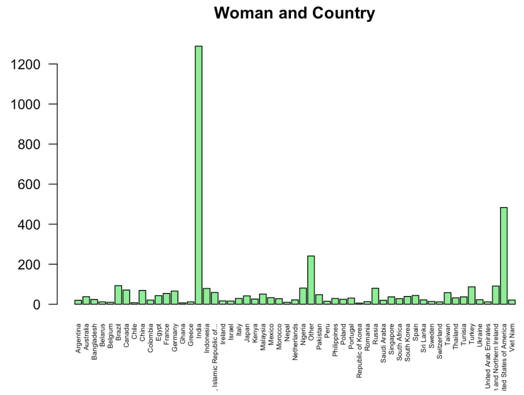
We are going to compare the men and women within the gender and majority counts they occupied compared the age. We will use separate graphs to show and analyze them individually then compare between them.

Analysis

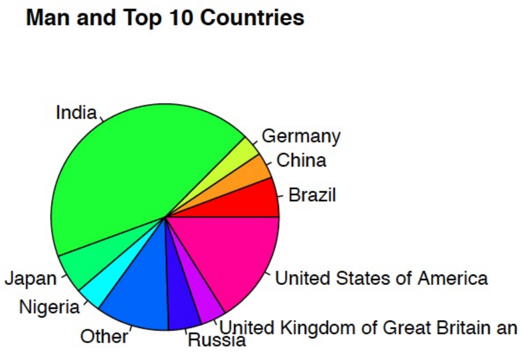


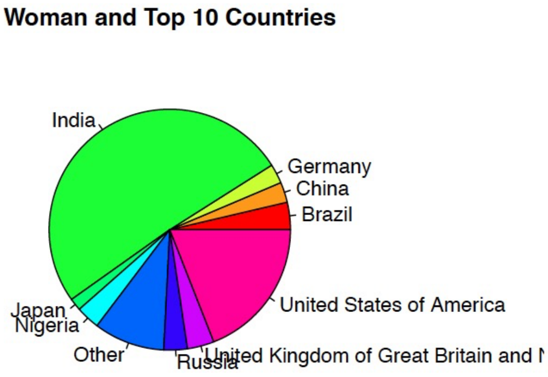


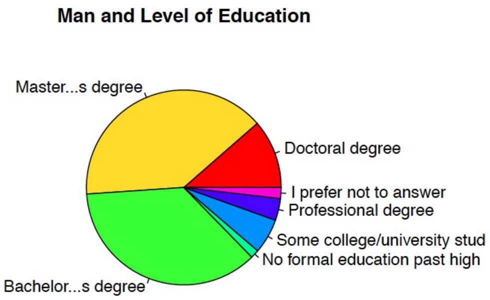


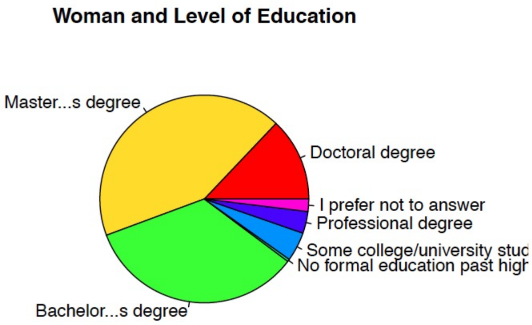


Country from left to right: Argentina, Bangladesh, Belgium, Canada, China, Egypt, Germany, Greece, Indonesia, Ireland, Italy, Kenya, Mexico, Nepal, Nigeria, Pakistan, Philippines, Portugal, Romania, Saudi Arabia, South Africa, Spain, Sweden, Taiwan, Tunisia, Ukraine, Id Northern Ireland, and Viet Nam.


Man and Top 10 Countries


Woman and Top 10 Countries


Man and Level of Education


Woman and Level of Education

The man and age pie graph shows that most people who are taking the survey are around 25-29 and when the age gets older there are less participants. From 18 years old to 29 years old from the man and age graph they are half of the response and the rest are another half of the participants. It could be concluded that most of the people in our participants are young people. Then we look at the women and age graph, the majority participants are around age 22-24 and from 18 years old to 29 years old are 3/4 of the participants and the 1/4 are from 30 to 70+. By comparison between two graphs, we can see that the women participants are slightly younger than the male participants. They both have fewer participants in age 70+, especially the women who are almost zero participants in age 70+. This is probably the reason that technology is not that often used by older people. Another look, we are looking at man and country. Country India has huge participants than all other countries. There are over 4,000 participants currently residing in India. States of American are the second biggest participant population. From the women and country graph it is similar that India has the most participants and the States of American are second large participants. Overall, the distribution are really similar. Top 10 countries are already significant so we are comparing the top 10 countries, men and women in the top 10 countries, the Brazi, China, Japan, and Russia catch the attention because in the man and top 10 countries graph, their areas look different. China has slightly more than Germany but China and Germany are around the same for women. In the man's graph, Brazil has almost half the population than

Germany. However, In woman's graph it is only slightly higher than Germany. The most different two countries are Japan and Russia. In women and top 10 countries, India women have half of them. Man and top 10 countries are less than half of them. Both of them show from the comparison that men have more population than women. Even the rank of them has changed. When we look at the education level separately we could see that the womens are having more master degrees than bachelor degrees and their ratio is slightly more than the males's master degree and bachelor's degree. The highest level of education is a master's degree. By comparing the percentage of women doctor degrees to man doctor education, the women who have doctoral degrees are slightly higher.

Conclusion

The show shows that most of the participants are young and most of them are less than 40 years old. The main participants are from the country India and Japan and Russia are having much more men in the participants then women. When we look at the education level, the women actually are more willing to have higher degrees than male.

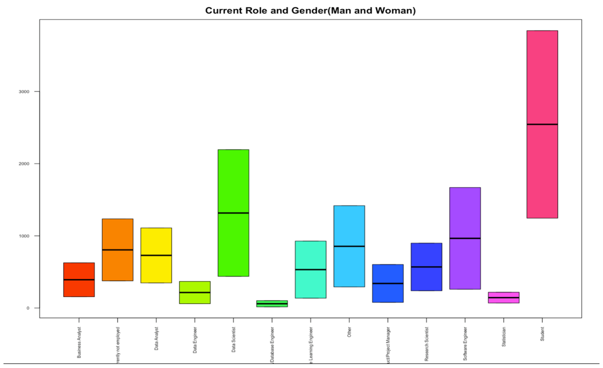
2.3 Working Experiences

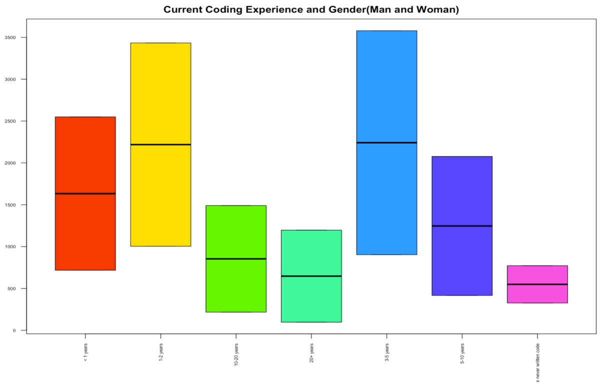
In this part, we will figure out the connection between Roles and Coding experiences, and make both point and interval estimates for the fraction of students with few coding experiences. Then check if gender affects this.

Method

We are using a boxplot to show their current role and their time of experience in coding. Then make a comparison of them and analyze the graph for the distribution of current role and the coding experience. Also find the estimator and CI in different cases.

Analysis





The upper bound of the box plot represents the male participants. The lower bound of the box plot represents the women participants. The line between the upper bound and lower bound is the average between the men and women participants. As we see the first graph shows the student has the largest difference between men and women. It is also the highest participants in their own gender to other roles. The data scientists are the second largest participants who are taking roles in. The least role in the graph is database Engineer which is really equally between men and women. Most people do not like to be database engineers. The software engineer is the third top participant who has a role in. Then look at the second graph. It shows the timeline of their coding experience. We see that the most women participants are experienced coding 1 to 2 years. For male, most of them have 3 to 5 years coding experiments. Refer back to the second question, we know that women experience coding at a younger age than male. It totally makes sense that women have fewer coding experience. Majority Male in the survey are a little bit older than women so they are the people who experience coding 3 to 5 years. Then we see that the male and female in different roles or years of experiment are really similar. As two biggest participants in two graphs, it could show that students have few coding experience. Therefore I am going to estimate whether gender really matters or not in the survey. There are too many variables in the survey. I am only using the new student to compare. I am comparing the new student with few coding experience(coding experiment less than 2 years) to all the students and all the participants. I found out that the estimator for comparing all the students is 0.5939 which means that almost 60% of the participants are students with less experience. The 95% confidence interval is from 0.59 to 0.61. By comparing all the participants, it shows the esmator equal to 0.153 and 95% confidence interval is 0.148 to 0.158. From this I am going to use the fraction with gender. I am going to see the male in students with few experiments compared to all the male students and also to all the participants. I got an estimate equal to 0.596 and 0.114 and their confidence interval is 0.58 to 0.6117 and 0.109 to 0.118. Then last I see the female with less experience compared to female students and all participants. I still got similar results which is 0.59 and Confidence interval 0.564 to 0.619. By comparing all the participants we got 0.036 and range from 0.034 to 0.039. By comparing those variables, we see that the new students with less experience compared to all students, new male students with less experience compared to tall male students, and the new female students with less experience compared to all the female students we got a really similar result for our estimator. Also the confidence intervals for three cases are really similar. We also know that when we add up male new students to female new students who both have less experience, we will get the answer equal to the first estimate case. In the calculation case, 0.114 + 0.036 is equal to 0.15. Therefore, we could conclude that gender does matter in this case but the others don't matter at all. Around 60 percent are the majority of new students who have less coding experience compared to all the students.

Conclusion

From the box plot, we can see that male and women have their most participants in the role of students. From the other box plot, it shows that men with 3 to 5 years coding experience are the majority and women with 1 to 2 years coding experience are most common. I checked three cases which are new students classified as students with less than 2 years coding experience. Then I make a comparison to all the students. I also did differences between male new students and female new students compared to their gender. I found out that the majority is 60 percent.It could be concluded that there does have differences between male and women but others do not matter.

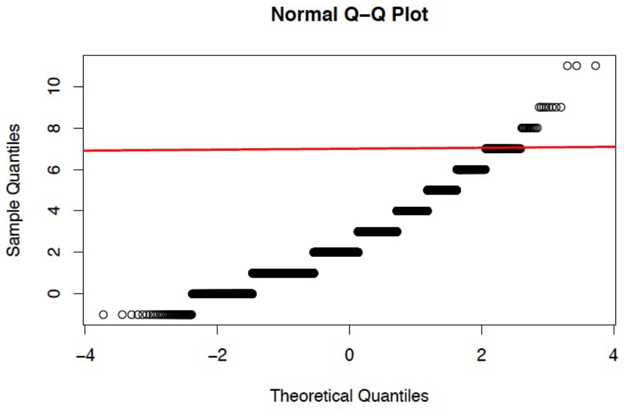
2.4 Languages

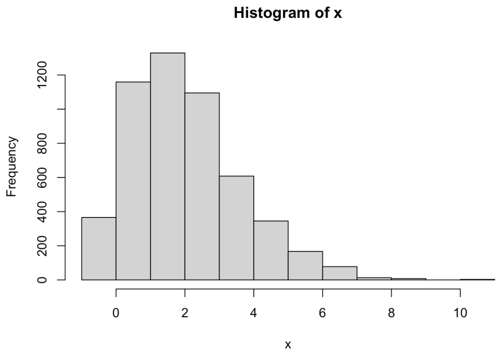
Coding languages are the most essential parts of data work. So in this part, we will make both point and interval estimates for the average number of languages used on a regular basis for new student coders.

Method

Using the sample statistic to see the difference between learning more languages. Does the longer time with coding experiment learning more languages than those who are new to coding. By using the Normal Q- Q plot to estimate the interval estimate.

Analysis





Mean	2.508606	
CI	2.463063	2.5541148
PI	-0.7663349	5.7835462

In R we calculated out the average number of known language estimated mean is 2.508 and the confidence interval is -0.766 to 5.783. The Normal Q-Q plot is roughly skewed to the right. We are comparing variables in the survey question 6 that ask for the year of writing code compared to the number of languages they are known. In a common sense, people who are more good at data science are those who know more languages than others. Then from the confidence interval, we could see that when there is 5 percent of significance level they would construct the confidence interval above and this means that they will cover the most of the true mean population. Average around 2 and 5 languages per person they have learned. From the graph we see the trend that as the sample quantities increase the theoretical quantiles will increase at the same time which means as the years of experience get longer they will learn more languages. From this stimulation study we see that the graph of histogram is shifted to the right.

Conclusion

By doing a simulation study on another language and year they experiment with coding. It was skewed to the right and they may not be satisfied when they have observed observations.Then most data are from the left side with tails. They do not have a strong correlation between the year of experience coding to the language they learned.

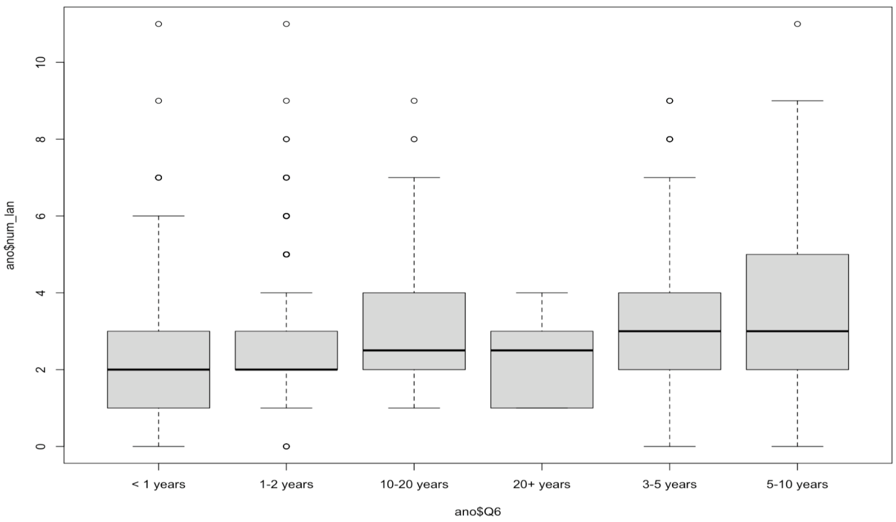
3 Advanced Analysis

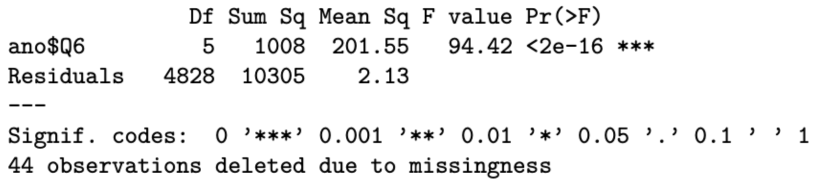
After the previous analysis we did above, there is a question: does the longer coding experience coders have, the more languages they usually use on a regular basis? Do the different coding experiences imply different numbers of languages they use? In this part, we will figure it out.

Method

In R, we could set a ANOVA test to check this question. Therefore, the Null hypothesis is that the average number of languages used are the same in every kind of experience. The alternative hypothesis is that some average numbers are different from others. Here we removed all the data that "I have never written code" due to it has all 0s and does not have any meaning here.

Analysis





```
              Df Sum Sq Mean Sq F value Pr(>F)    
ano$Q6         5   1008  201.55   94.42 <2e-16 ***
Residuals   4828  10305    2.13                   
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
44 observations deleted due to missingness
```

In R, first we made the boxplot. Having the general glance of the means shown as black horizontal lines in each box, it seems roughly that the means are relatively similar. Which in favor of the Null hypothesis. But after we did the ANOVA test seriously in R, we found that the p-value is smaller than 2e-16 which is almost 0. Such small p-values induce the rejection of the Null hypothesis in favor of alternative hypotheses. That these are some average numbers of languages they use on a regular basis different from others. The means are not all the same.

Conclusion

ANOVA test is quite a powerful tool which generalizes many of the frameworks, and tests all of them at once. Applying the ANOVA test to our data, we could conclude the average number of languages used on a regular basis is not the same in every time group.This answers the question and tells us that there is no clear correlation between number of languages used and coding experiences.

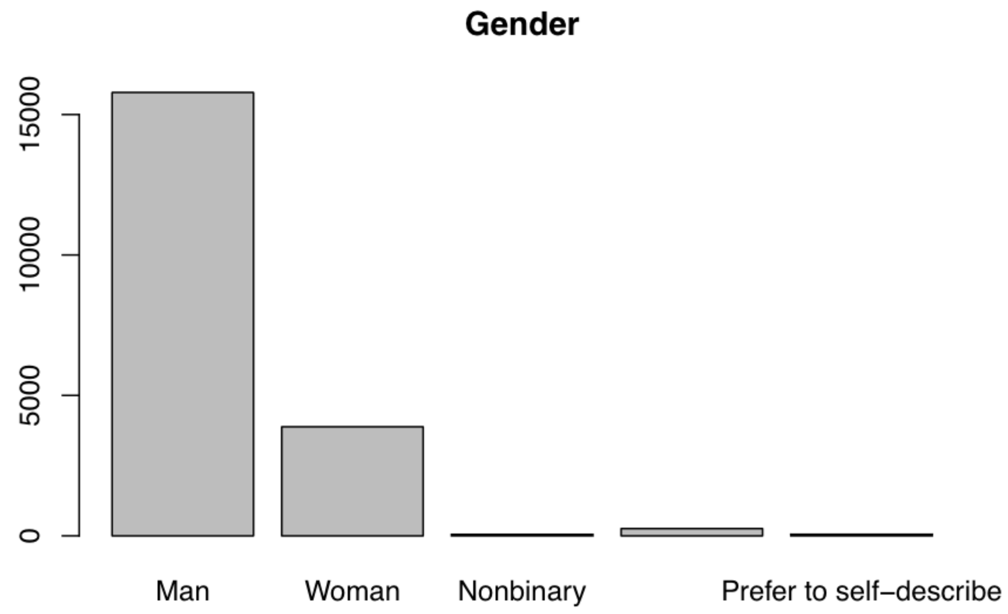
4 Conclusion and Discussion

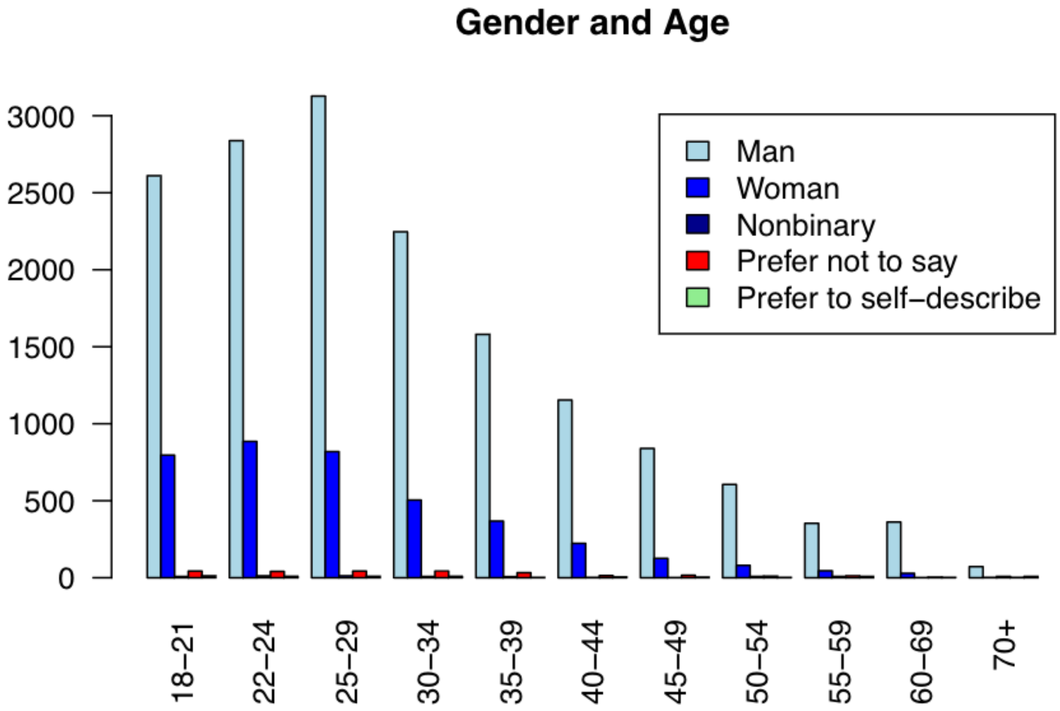
In this analysis, we are using a survey that is trying to make a comprehensive view of the different country’s marching learning and data science. From 2017 to 2019 this survey took place and half weeks in October. We are given 20,035. final valuable responses. We have multiple choice questions in this survey. Those questions could be what is their highest level of formal education, how many years have they been writing code.

Data set included 20,035 observations and 355 columns. We are solving to see whether the stereotype for gender is true or not. This stereotype is related to the data area. Do we have more male than females in the data area. I checked the background of the dataset and found out there are 4 times larger males participants than female and then I compared their age, country , and education level. We obviously see that there are more male in each category than females. Going forward, I checked the differences between women and male in different categories by using proportions. This gives me that they are really similar, only differences are really slight. Such as the education level, the women more willing to have higher degrees than male and country wsie Japan are having more male participants than female a lot. In question 3 I actually found out that a majority is 60 percent between new students with less experience to all the students. And also using the gender to be the variable to see the difference. But it is actually still the same so it can be concluded that there does have differences between male and women but others do not matter. By using q-q plot I found out that there is a biased relationship between the year of experience coding to the language they learned. Last, using the ANOVA test, we find out that the average number of languages used on a regular basis is not the same in every time group.This tells us that there is no clear correlation between number of languages used and coding experiences.

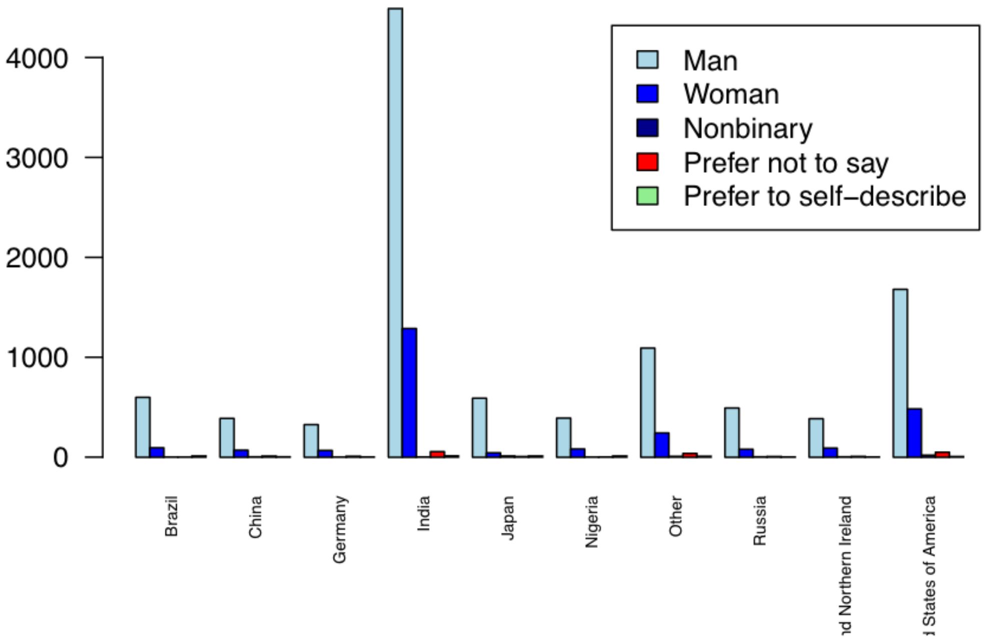
Appendix

Full view of Graphs used in 2.1

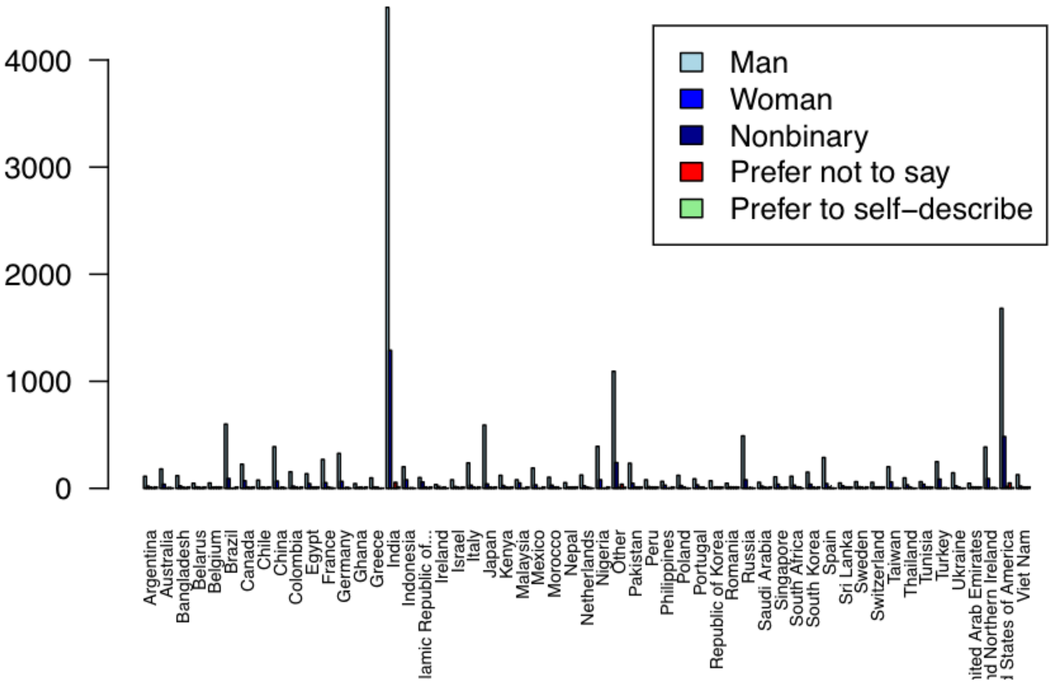


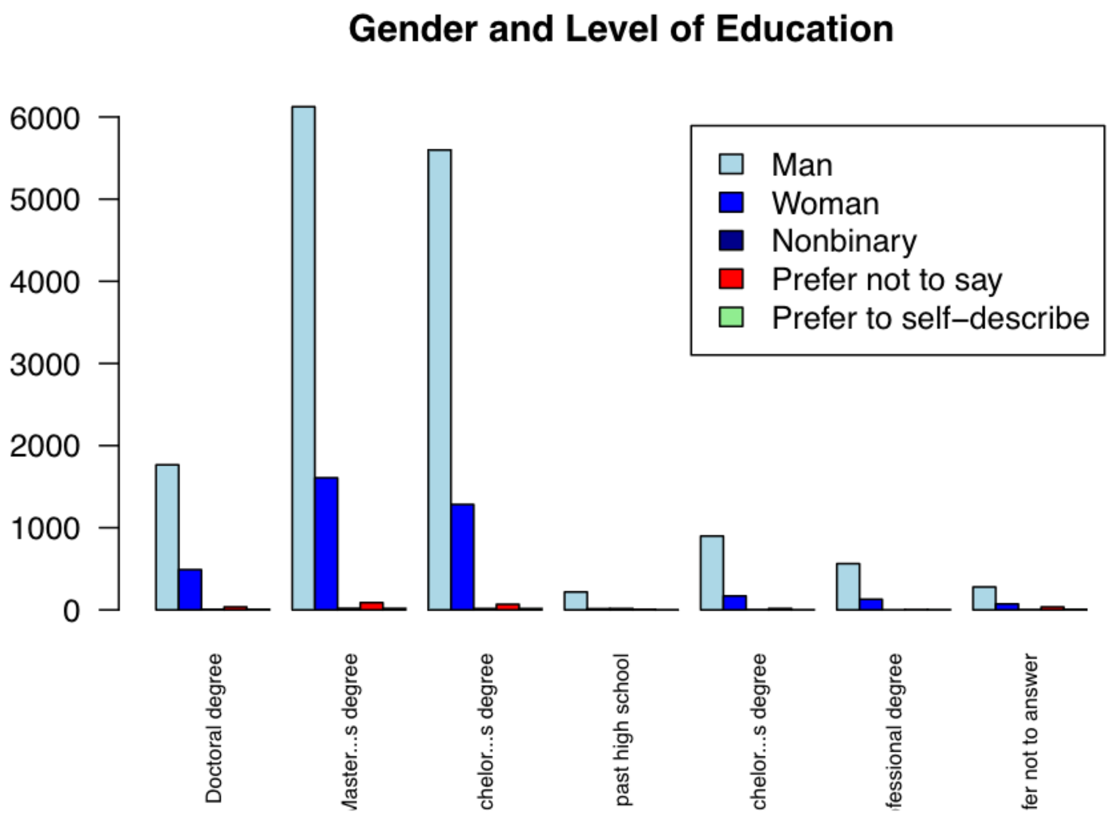


Gender and Top 10 Countries



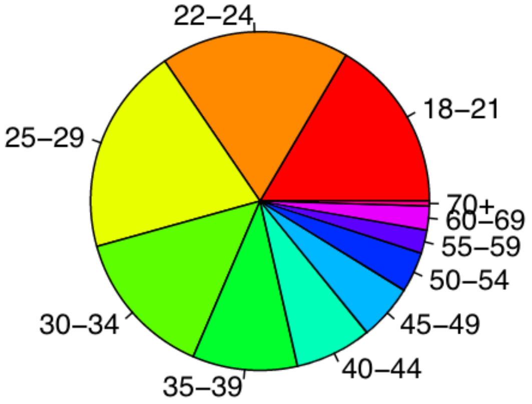
Gender and Country



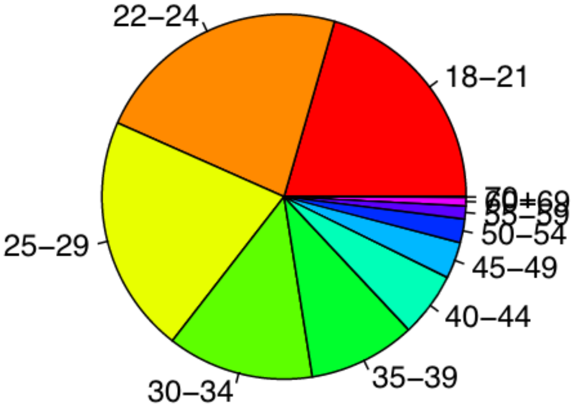

Gender and Level of Education
6000
5000
4000
3000
2000
1000
0
Man
Woman
Nonbinary
Prefer not to say
Prefer to self−describe
Doctoral degree
Master...s degree
chelor...s degree
past high school
chelor...s degree
fessional degree
fer not to answer

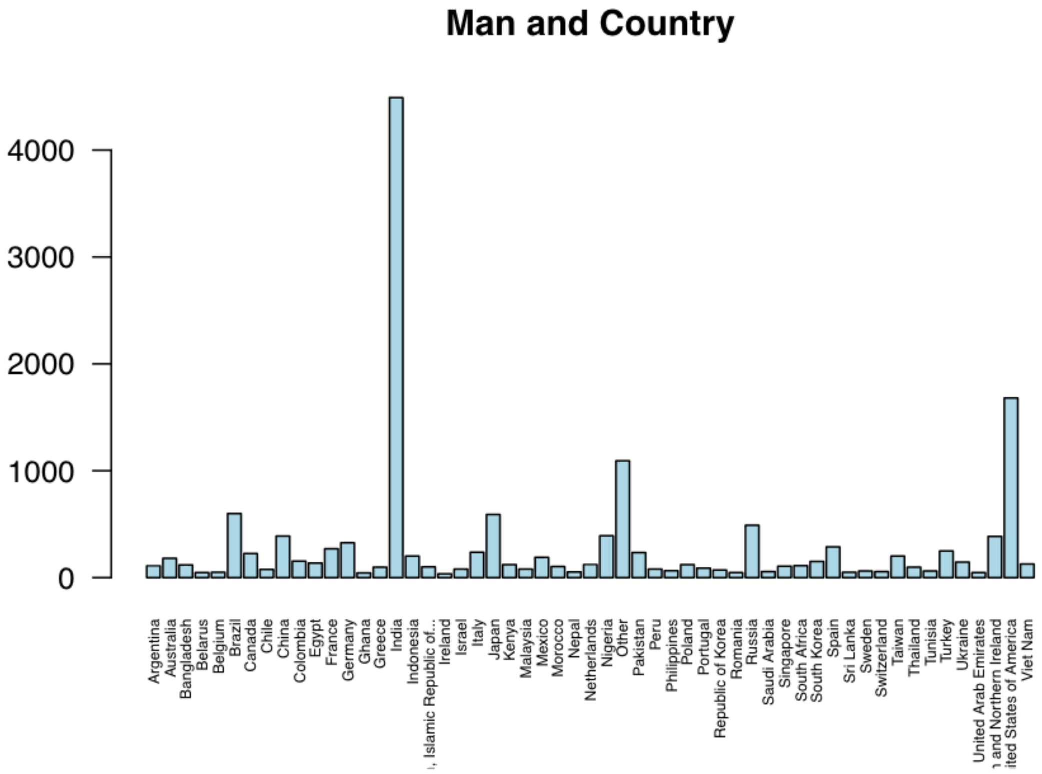
Full view graphs used in 2.2

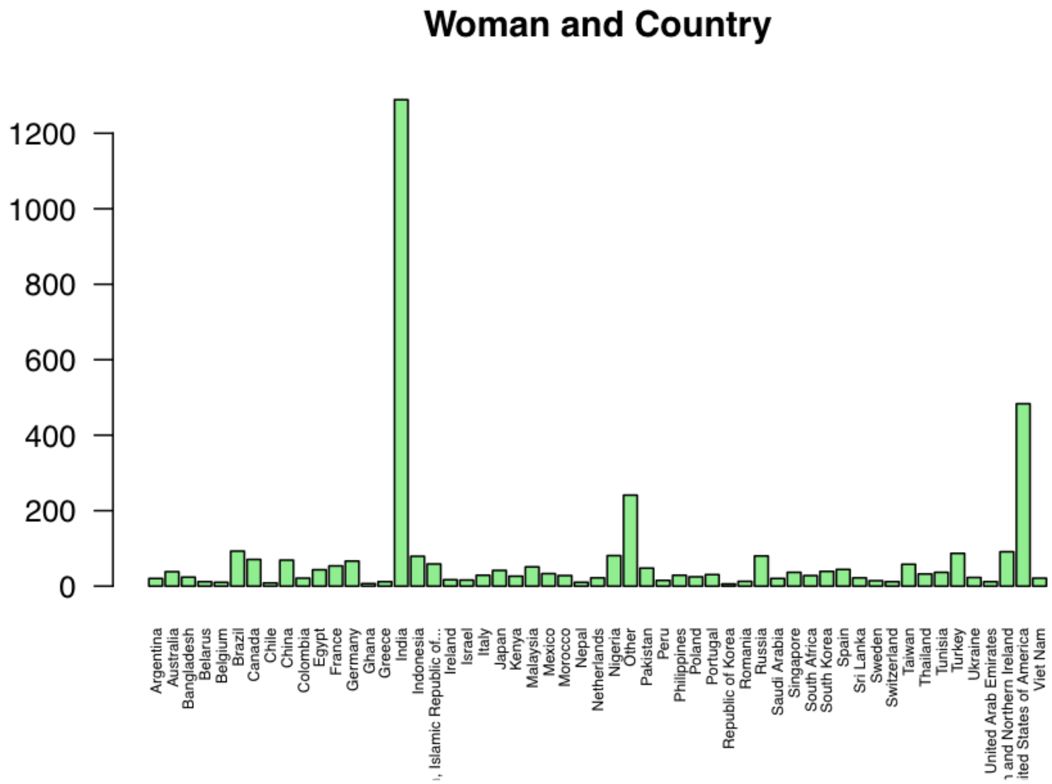
Man and Age



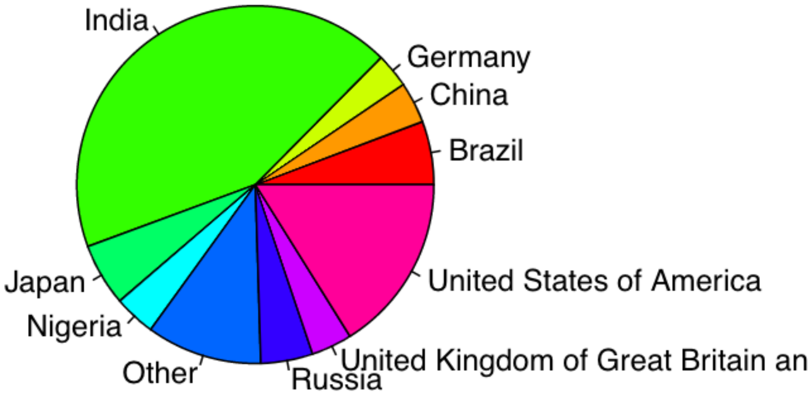
Woman and Age



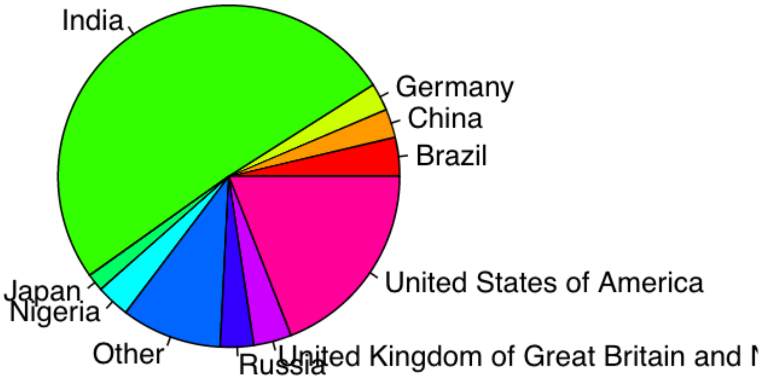

Man and Country
0
1000
2000
3000
4000
Argentina
Australia
Bangladesh
Belarus
Belgium
Brazil
Canada
Chile
China
Colombia
Egypt
France
Germany
Ghana
Greece
India
Indonesia
ı, Islamic Republic of...
Ireland
Israel
Italy
Japan
Kenya
Malaysia
Mexico
Morocco
Nepal
Netherlands
Nigeria
Other
Pakistan
Peru
Philippines
Poland
Portugal
Republic of Korea
Romania
Russia
Saudi Arabia
Singapore
South Africa
South Korea
Spain
Sri Lanka
Sweden
Switzerland
Taiwan
Thailand
Tunisia
Turkey
Ukraine
United Arab Emirates
ı and Northern Ireland
ited States of America
Viet Nam


Woman and Country
0
200
400
600
800
1000
1200
Argentina
Australia
Bangladesh
Belarus
Belgium
Brazil
Canada
Chile
China
Colombia
Egypt
France
Germany
Ghana
Greece
India
Indonesia
ı, Islamic Republic of...
Ireland
Israel
Italy
Japan
Kenya
Malaysia
Mexico
Morocco
Nepal
Netherlands
Nigeria
Other
Pakistan
Peru
Philippines
Poland
Portugal
Republic of Korea
Romania
Russia
Saudi Arabia
Singapore
South Africa
South Korea
Spain
Sri Lanka
Sweden
Switzerland
Taiwan
Thailand
Tunisia
Turkey
Ukraine
United Arab Emirates
ı and Northern Ireland
ited States of America
Viet Nam

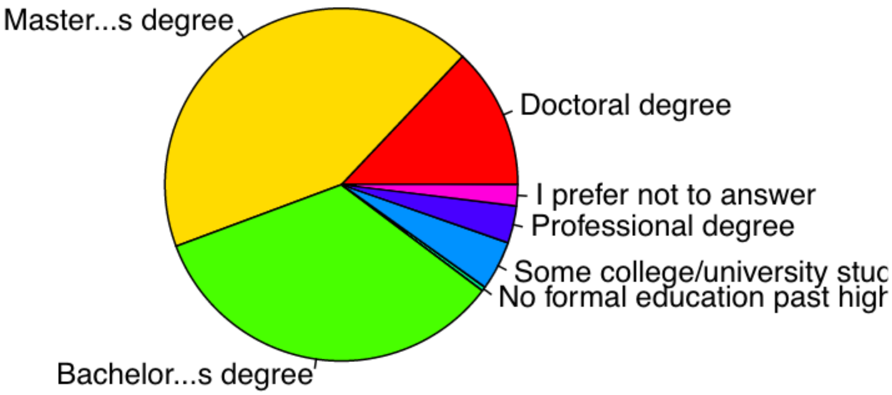
Man and Top 10 Countries



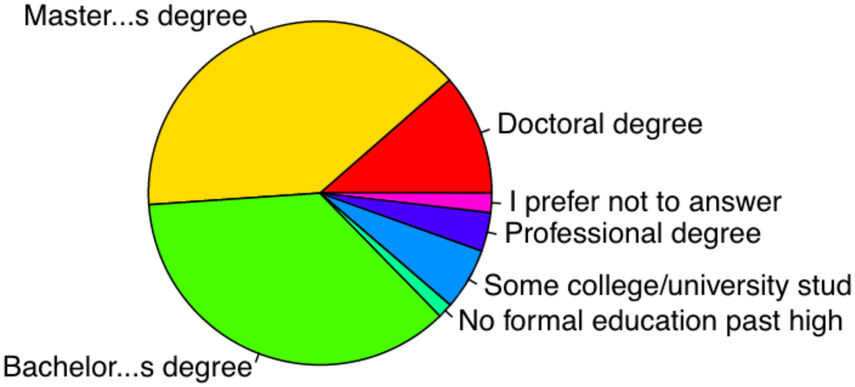
Woman and Top 10 Countries



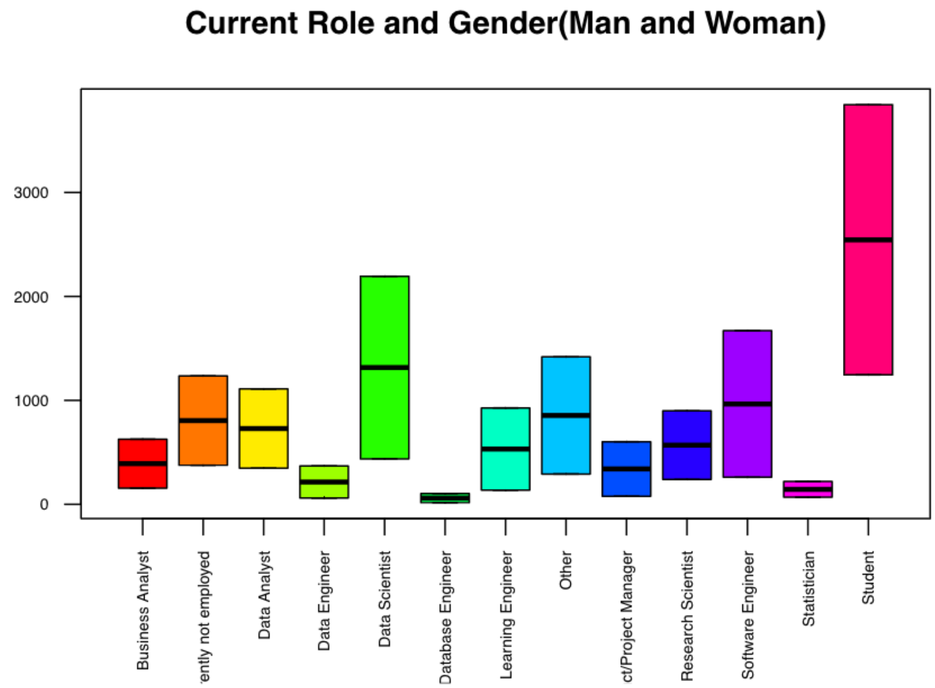
Woman and Level of Education



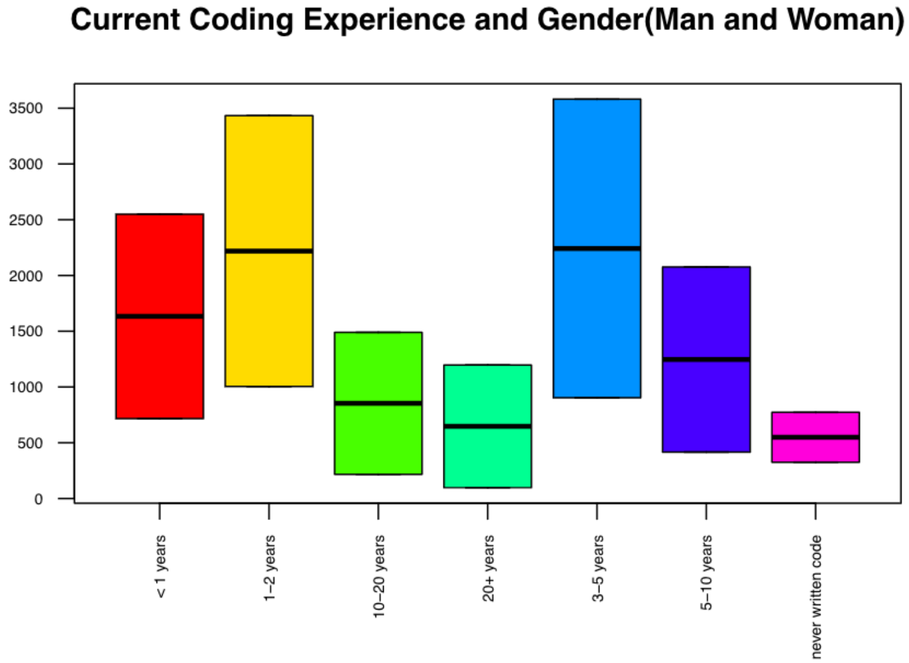
Man and Level of Education



Full view graphs used in 2.3

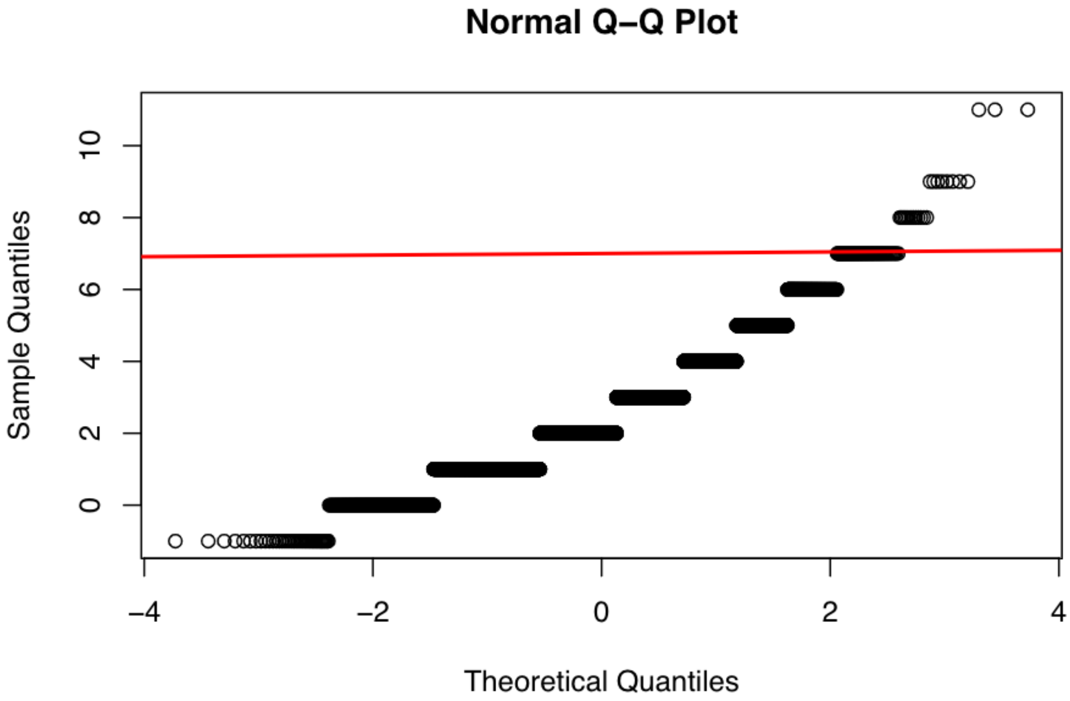


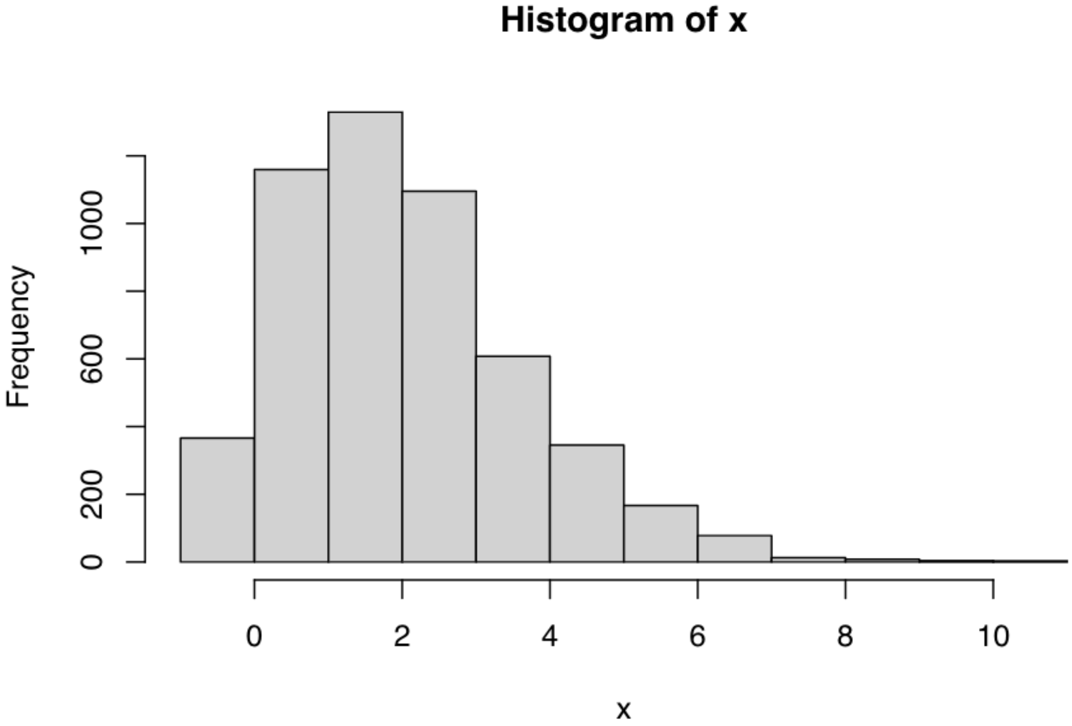
Current Role and Gender(Man and Woman)



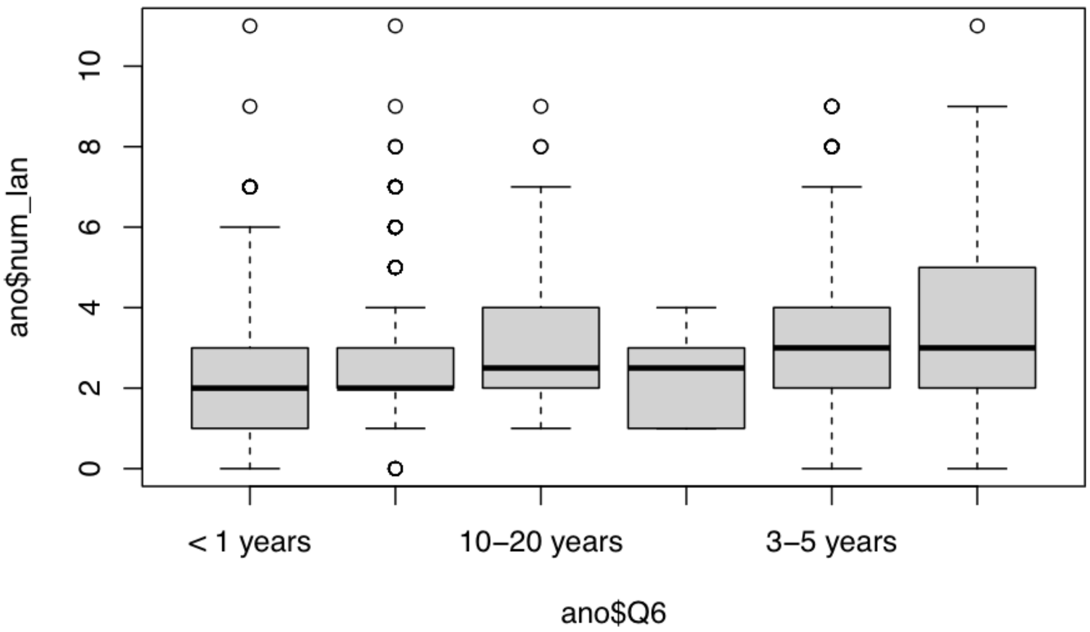
Current Coding Experience and Gender(Man and Woman)

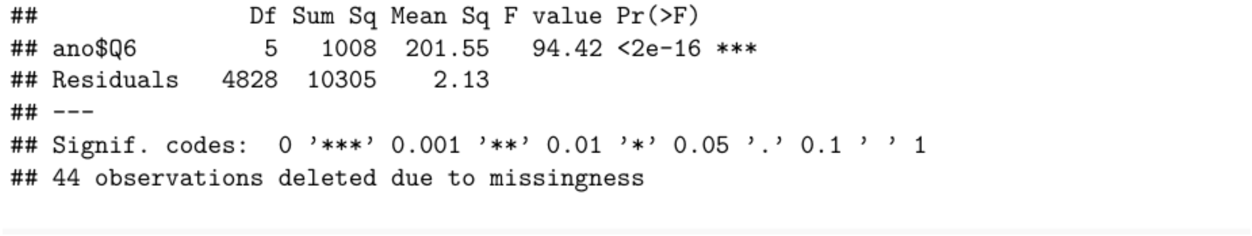
Full view graphs used in 2.4





Full view graphs used in 3





```
##                Df Sum Sq Mean Sq F value Pr(>F)    
## ano$Q6          5   1008  201.55   94.42 <2e-16 ***
## Residuals    4828  10305    2.13                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 44 observations deleted due to missingness
```