Jerry Cheng

Dr. Bixler

Computer Science with Data Analysis

30 January 2018

Computer Science with Data Analysis Final Project

**Introduction:**

Stack Overflow is one of the largest online communities dedicated for the education, sharing of knowledge, and construction of careers pertaining to the field of computer sciences. With over 50 million monthly visitors, an average of 51,000 concurrent users, and over 14 million questions and answers, Stack Overflow is a trusted tool amongst programmers and developers throughout the world. In the month of January in 2017, Stack Overflow conducted a survey, questioning various aspects about the userbase of the website. Various questions were given to users who took the survey, varying from background details (including race and gender), to even more trivial questions (such as how to properly pronunciation for the acronym for graphics interchange format, GIF). This project highlights some of the data from this survey, providing a graphic visualization in the form of a multitude of graphs. It also serves to provide an answer to these survey questions in a clean, visual point of view, without the bulkiness of a large spread sheet. These graphs were constructed using the programming language R, a language created for statistical analysis using the program RStudio. The data used for this project originated from the dataset sharing website Kaggle.com, who in turn obtained this data directly from Stack Overflow.

**Pre-Graph Code\*:**

*\*The code here does not detail any starting processes that must be accomplished before the utilization of this code. Before using this code, R and RStudio must be installed, and the dataset must be obtained*.

Note: all the code here is made under the assumption that this chunk of code was done prior. This code will install the packages necessary for creating the graphs.

Install.packages(“tidyverse”)

Install.packages(“ggthemes”)

Library(“ggthemes”)

Library(“tidyverse”)

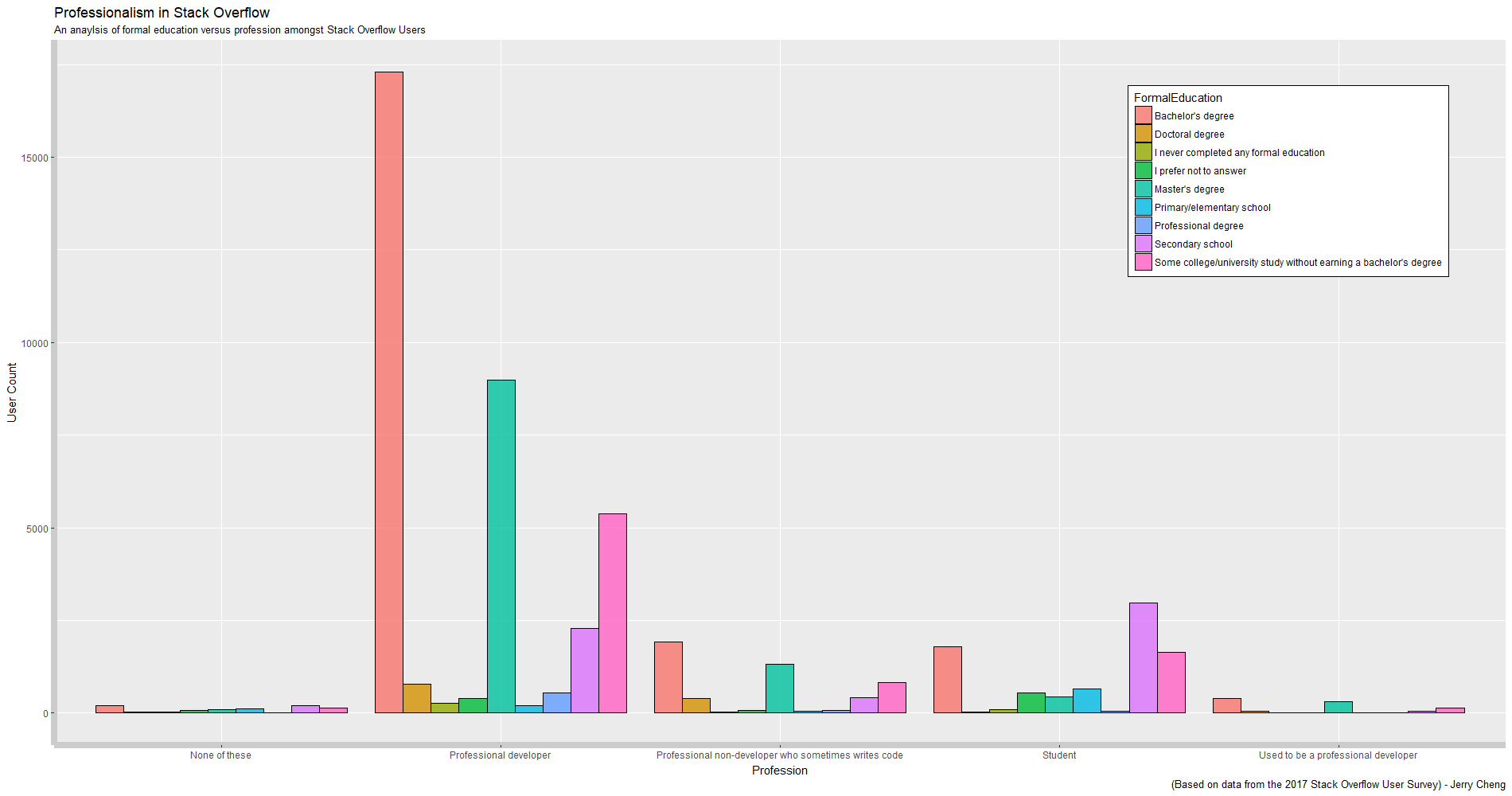
The dataset must also be uploaded to RStudio through the navigation of the GUI. In this case, the data set used will be referred to as survey\_results\_public.

**Data Excerpt:**

|  |  |
| --- | --- |
| Respondent | Professional |
| 1 | Student |
| 2 | Student |
| 3 | Professional developer |
| 4 | Professional non-developer who sometimes writes code |
| 5 | Professional developer |
| 6 | Student |
| 7 | Professional non-developer who sometimes writes code |

*A random excerpt of data to detail the format in which the data is cast in*

**Graph 1:** Are users on Stack Overflow educated? And what profession do they hold?



Block: ggplot(data = survey\_results\_public, aes(x = Professional, fill = FormalEducation)) + geom\_histogram(stat = "count", col = "black", alpha = 0.7, position = "dodge") + labs(title="Professionalism in Stack Overflow", subtitle = "An anaylsis of formal education versus profession amongst Stack Overflow Users", caption = "(Based on data from the 2017 Stack Overflow User Survey) - Jerry Cheng") + labs(x = "Profession", y = "User Count") + theme(legend.background = element\_rect(fill = "white", size = 0.5, linetype = "solid", color = "black")) + theme(legend.position = c(.85, .8)) + theme(axis.line = element\_line(size = 3, color = "grey80"))

Graph Analysis:

This graph compares different factors of the Stack Overflow userbase, notably comparing whether the user is a professional developer versus the highest level of formal education they have obtained, versus user count. Many of the users who have taken the Stack Overflow survey consider themselves to be professional developers. The majority of these professional developers have obtained a Bachelor’s degree, followed by the number of users who have obtained a Master’s degree, followed by those who have completed some of college. The next highest group of users identify as being currently students, with the majority of them in the process of obtaining an undergraduate degree.

Code Analysis:

1. ggplot(data = survey\_results\_public, aes(x = Professional, fill = FormalEducation)) +
2. geom\_histogram(stat = "count", col = "black", alpha = 0.7, position = "dodge") +
3. labs(title="Professionalism in Stack Overflow", subtitle = "An anaylsis of formal education versus profession amongst Stack Overflow Users", caption = "(Based on data from the 2017 Stack Overflow User Survey) - Jerry Cheng") +
4. labs(x = "Profession", y = "User Count") +
5. theme(legend.background = element\_rect(fill = "white", size = 0.5, linetype = "solid", color = "black")) +
6. theme(legend.position = c(.85, .8)) +
7. theme(axis.line = element\_line(size = 3, color = "grey80"))

Line 1 tells the computer to use ggplot, used for graphing plots. It tells the computer that the data that must be used should be obtained from the dataset named “survey\_results\_public”. The next part of the line tells the computer that the x value should from the “Professional” column, and the filled in colors of the graph should be form the “FormalEducation” column.

Line 2 tells the computer that the desired graph to be created is a histogram, through “geom\_histogram”. The next piece , “stat = “count”” tells the computer that the statistic for the y axis that should be used is the frequency, or count. It is **imperative** that this line is used, as the format the data is in requires it. “col = “black”” will set a black outline around the bars in the histogram. “alpha = 0.7” changes the opacity of each bar. “position = “dodge”” will separate individual values in each individual bars, helping to clarify the data.

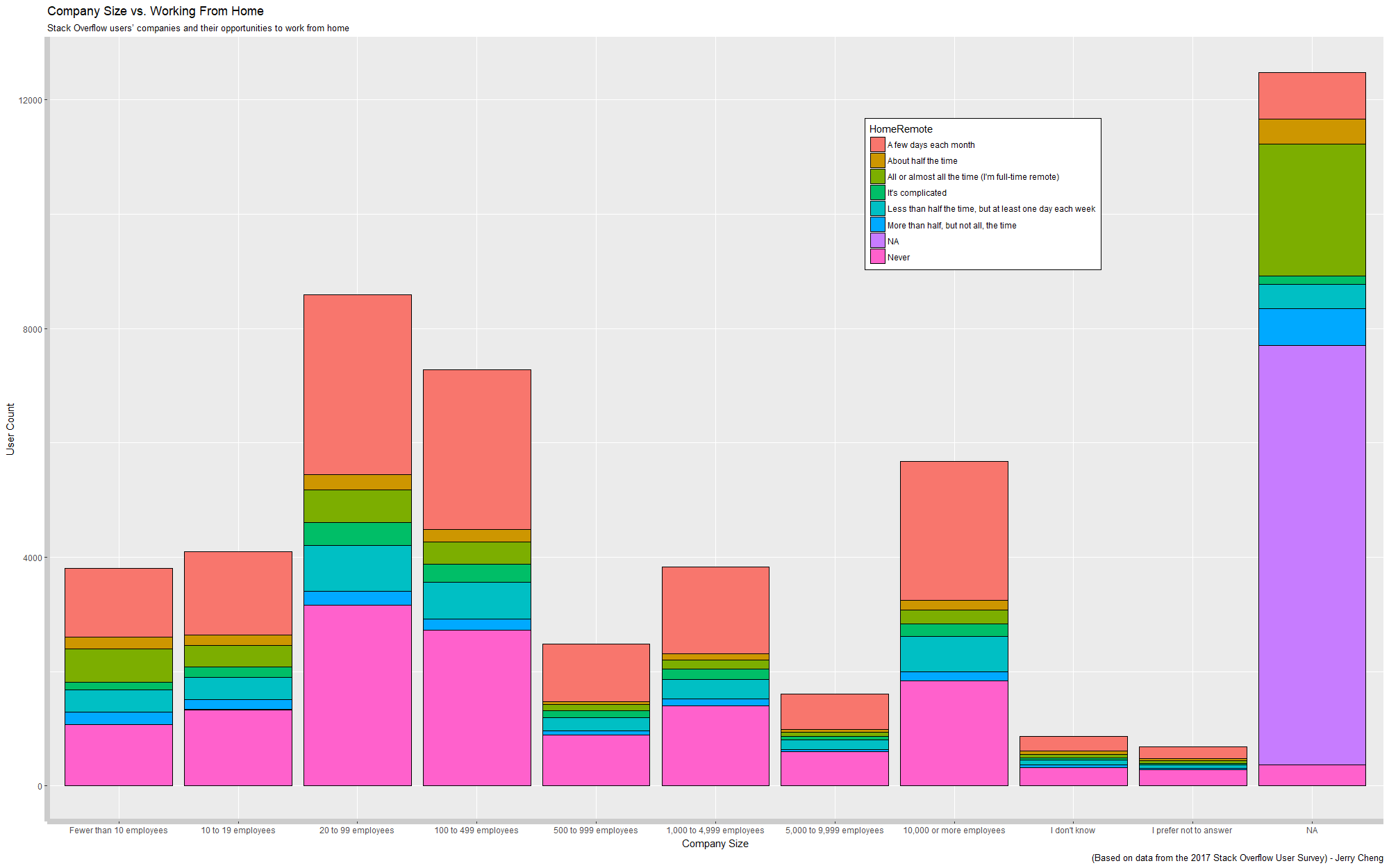
Line 3 customizes some of the labels on the graph, specifically the title, subtitle, and caption. This serves purely for aesthetic purposes.

Line 4 customizes the label titles on the x and y axes.

Line 5 customizes the legend, notably the fill color, the border color, the border line type, and the size of the legend.

Line 6 customizes the position of the legend, x location by y location. This line serves two purposes: by moving the legend position, it will keep intact an ideal aspect ratio for the graph itself, as well as preventing it from blocking data. It is important to make sure the legend is in an ideal location. If it is located outside of the plot, it will mess with the aspect ratio.

Line 7 adds grey lines, size 3, on the two axes of the graph. It serves purely for aesthetic purposes.

**Graph 2:** How large are the companies Stack Overflow users work for, and how often do they get to work from home? 

Block 1: survey\_results\_public$CompanySize <- factor(survey\_results\_public$CompanySize, levels = c("Fewer than 10 employees", "10 to 19 employees", "20 to 99 employees", "100 to 499 employees", "500 to 999 employees", "1,000 to 4,999 employees", "5,000 to 9,999 employees", "10,000 or more employees", "I don't know", "I prefer not to answer", "NA"))

Block 2: ggplot(data = survey\_results\_public, aes(x = CompanySize, fill = HomeRemote)) + geom\_bar(col = "black") + theme(legend.position = c(.7, .8)) + theme(legend.background = element\_rect(size = 0.5, linetype = "solid", color = "black")) + labs(title = "Company Size vs. Working From Home", subtitle = "Stack Overflow users’ companies and their opportunities to work from home", caption = "(Based on data from the 2017 Stack Overflow User Survey) - Jerry Cheng") + labs(x = "Company Size", y = "User Count") + theme(axis.line = element\_line(size = 3, color = "grey80"))

Graph Analysis:

This graph compares, out of the Stack Overflow userbase who work for a company, how large the companies they work for are, and how often they are able to work from home. For the majority of the users who answered, they chose to skip this question. Ignoring this portion of the data, it seems that the majority of users who answered this question work in a small company in the size range of 29 to 99 employees. And the about 2/3 of these people who work in this company size will either never get to work from home, or only get the opportunity to work from home a couple of days per month. This 2/3 ratio seems to be consistent amongst most company sizes, excluding companies within the size range between 5,000 and 9,999 employees. It seems that if one wanted to work from home a lot, it would be advisable for them to avoid companies within that size range.

Code Analysis (Block 1):

Because of the irregular type of data the dataset in presented in, it is necessary to manually set the levels for each of the data points. Without doing so, R will organize the data in an undesirable order. This block of cold essentially just reorders the factor levels for the CompanySize data.

Code Analysis (Block 2):

1. ggplot(data = survey\_results\_public, aes(x = CompanySize, fill = HomeRemote)) +
2. geom\_bar(col = "black") +
3. theme(legend.position = c(.7, .8)) +
4. theme(legend.background = element\_rect(size = 0.5, linetype = "solid", color = "black")) +
5. labs(title = "Company Size vs. Working From Home", subtitle = "Stack Overflow users’ companies and their opportunities to work from home", caption = "(Based on data from the 2017 Stack Overflow User Survey) - Jerry Cheng") +
6. labs(x = "Company Size", y = "User Count") +
7. theme(axis.line = element\_line(size = 3, color = "grey80"))

Line 1 calls upon ggplot, telling the computer that a graph will be printed out. It tells the computer that the data should come from the survey\_results\_public spreadsheet, the x values should come from the CompanySize data set, and the graph should be filled in accordingly to the HomeRemote data set.

Line 2 tells the computer that a bar graph should be made, and the outline to the individual bars. Because the code asks for a bar graph, it will default to using “count” for the y-values, discarding the need for the “stat = “count”” that was necessary for geom\_histogram.

Line 3 adjusts the position of the legend.

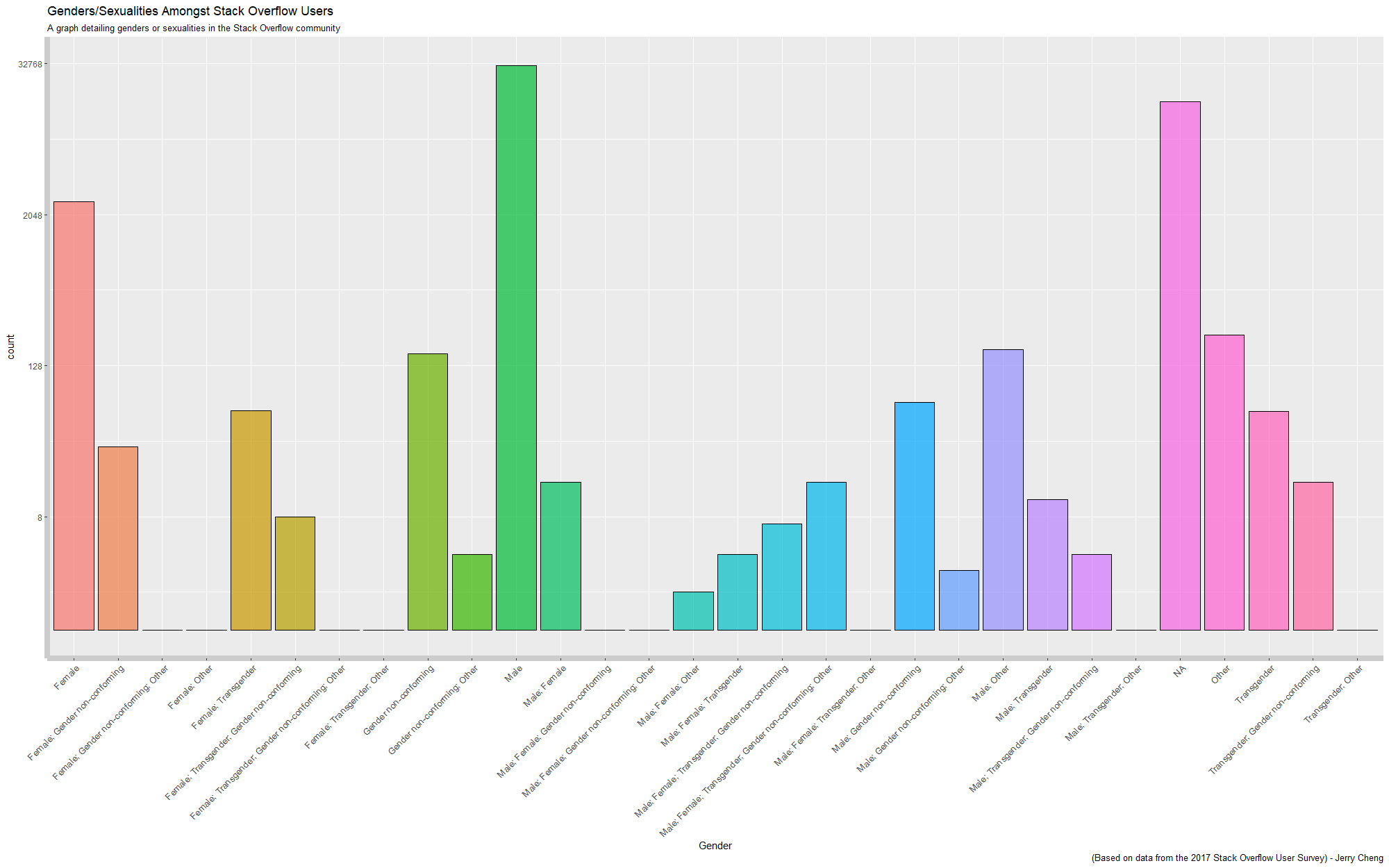
Line 4 customizes the legend – the size, type of border, and color of the border for the legend.

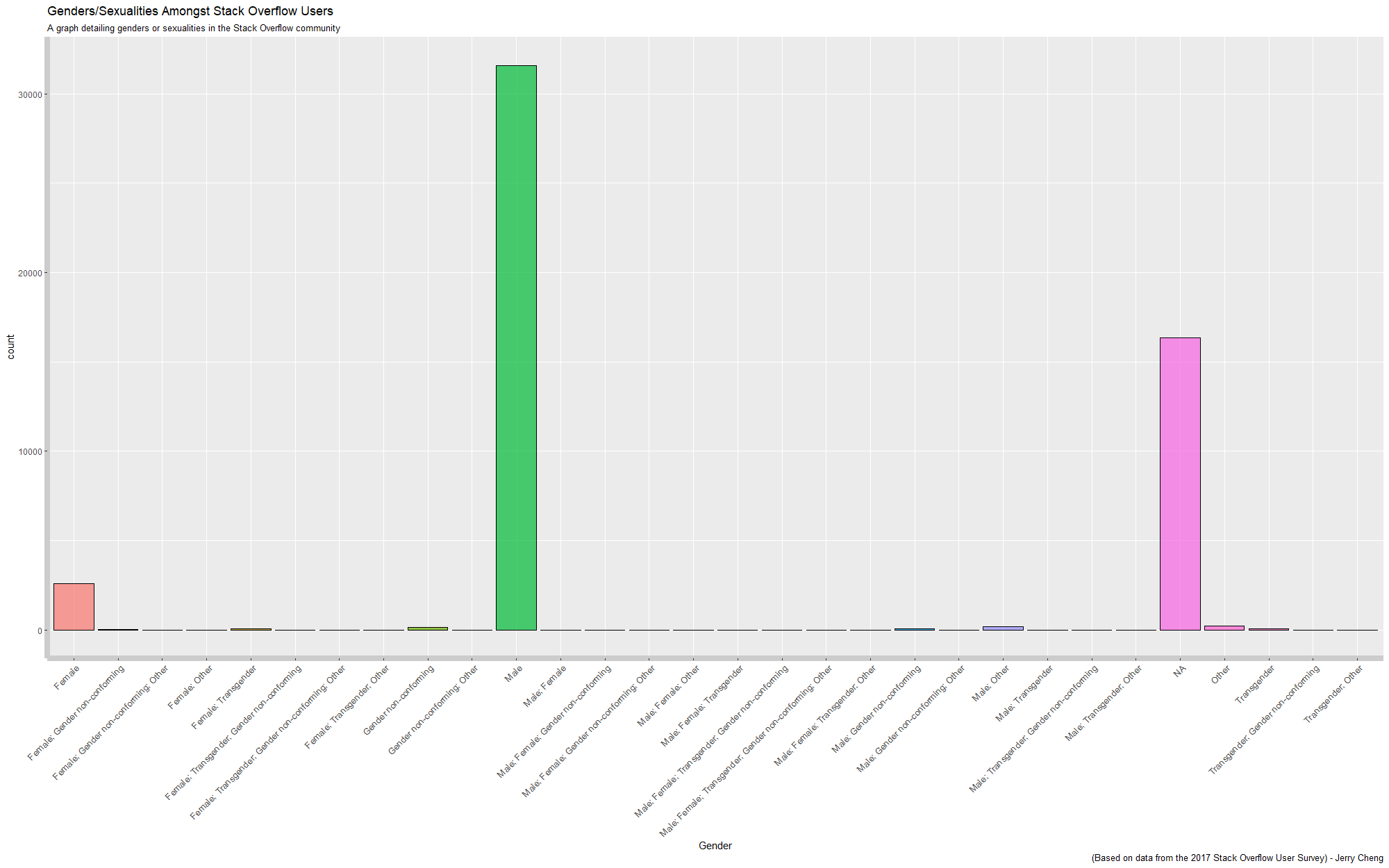
Line 5 customizes the labels for the graph, including the title, subtitle, and caption.

Line 6 customizes the labels for each of the axes.

Line 7 adds a shade of grey to each of the axes.

**Graph 3:** Is there a gender gap amongst Stack Overflow users, and if so, how large is it?





Block A: ggplot(data = survey\_results\_public, aes(x = Gender, fill = Gender)) + geom\_bar(col = "black", alpha = 0.7) + theme(axis.text.x=element\_text(size = 10, angle = 45, hjust = 1, vjust = 1)) + scale\_y\_continuous(trans = 'log2') + labs(title="Genders/Sexualities Amongst Stack Overflow Users", subtitle = "A graph detailing genders or sexualities in the Stack Overflow community", caption = '(Based on data from the 2017 Stack Overflow User Survey) - Jerry Cheng') + theme(axis.line = element\_line(size = 3, color = "grey80")) + theme(legend.position = "none")

Block B: ggplot(data = survey\_results\_public, aes(x = Gender, fill = Gender)) + geom\_bar(col = "black", alpha = 0.7) + theme(axis.text.x=element\_text(size = 10, angle = 45, hjust = 1, vjust = 1)) + labs(title="Genders/Sexualities Amongst Stack Overflow Users", subtitle = "A graph detailing genders or sexualities in the Stack Overflow community", caption = '(Based on data from the 2017 Stack Overflow User Survey) - Jerry Cheng') + theme(axis.line = element\_line(size = 3, color = "grey80")) + theme(legend.position = "none")

Graph Analysis:

These graphs compares the variety of different genders or sexualities amongst Stack Overflow users. Both graphs use the same exact data set. However, two graphs were made because of the necessity for comparison. The first graph, graph A, uses a logarithmic y axis in order to fully display all of the data. Graph B uses a linear y axis to show the true extent of the data differences. While graph A may be able to show all of the data without disregarding the handful of people who chose different options, graph B shows a more accurate representation of how many people voted for what. The results are staggering. If we were to disregard the N/A data, meaning that people skipped the question, there is an **overwhelming majority**, **over** **10 times** as many males as there are females on Stack Overflow. And if we were to add the N/A bar to the female bar, the male bar would **still be significantly** higher. Odds are, if you come across someone on Stack Overflow, they are probably a male. Whether or not Stack Overflow is an accurate representation of a gender gap in the computer science field remains to be seen- however, the data shows that there are a significantly higher male to female ratio on this website.

Code Analysis\*:

\**Because both graphs code-wise are nearly identical, I will only go in-depth with one of them. I will point out what makes the two different*.

1. ggplot(data = survey\_results\_public, aes(x = Gender, fill = Gender)) +
2. geom\_bar(col = "black", alpha = 0.7) +
3. theme(axis.text.x=element\_text(size = 10, angle = 45, hjust = 1, vjust = 1))
4. scale\_y\_continuous(trans = 'log2') +
5. labs(title="Genders/Sexualities Amongst Stack Overflow Users", subtitle = "A graph detailing genders or sexualities in the Stack Overflow community", caption = '(Based on data from the 2017 Stack Overflow User Survey) - Jerry Cheng') +
6. theme(axis.line = element\_line(size = 3, color = "grey80")) +
7. theme(legend.position = "none")

Line 1 tells the computer that ggplot needs to be used, and a graph will be made. It tells the computer that the data set that needs to be used is survey\_results\_public, the x value will be from the “Gender” column, and the colors should filled by the “Gender” data set.

Line 2 tells the computer that a bar graph should be made, with a black outline and an opacity of 0.7.

Line 3 adjusts the angle of the x axis labels. Because the labels for the different Gender groups are so long, leaving them horizontal results in the overlap of the many different labels. Therefore, it is necessary to rotate each of them by 45 degrees. The code specifically says that the labels should be of size 10, at an angle of 45, and that horizonal/vertical adjustments should be done with a factor of 1.

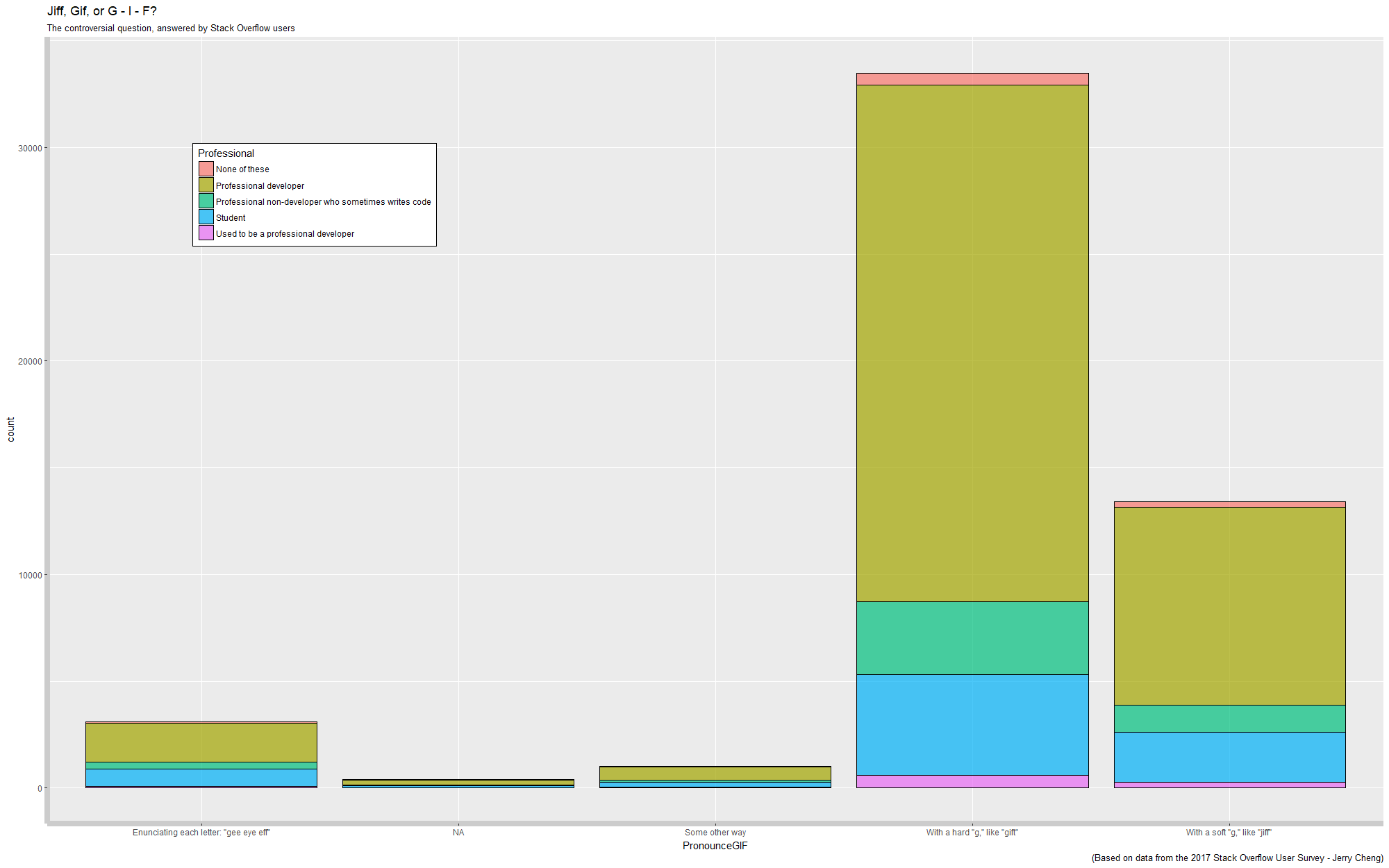
**Line 4** is the line that sets the two different graphs apart. Graph A has this line, graph B does not. This line adjusts the y axis scale logarithmically. It essentially causes the y axis to increase exponentially – slow at first, but rapidly scaling. Therefore, it is possible to see data from x values that have low values, as well as data that has high values.

Line 5 customizes the labels on the graph, including the title, subtitle, and caption.

Line 6 adds grey lines on each of the axes.

Line 7 eliminates the legend. Because a legend would be redundant on this graph, this line is used to simply remove the legend.

**Graph 4:** Jiff, Gif, or G – I – F? And who are they to say it?



Block: ggplot(data = survey\_results\_public, aes(x=PronounceGIF, fill = Professional)) + geom\_bar(col = "black", alpha = 0.7) + labs(title="Jiff, Gif, or G - I - F?", subtitle = "The controversial question, answered by Stack Overflow users", caption = "(Based on data from the 2017 Stack Overflow User Survey - Jerry Cheng)") + theme(legend.position = c(.2, .8)) + theme(legend.background = element\_rect(fill = "white", size = 0.5, linetype = "solid", color = "black")) + theme(axis.line = element\_line(size = 3, color = "grey80"))

Graph Analysis:

One of the most controversial questions amongst computer users everywhere. How do you pronounce GIF? Gee – Eye – Eff? Jiff? Or Gif? It seems that amongst the Stack Overflow community, GIF is most commonly pronounced with a hard “G”. For every different level of professional development, every demographic holds a majority in the hard “G” section. There are more professional developers that say GIF with a hard “G” than there are every other demographic combined.

Code Anaylsis:

1. ggplot(data = survey\_results\_public, aes(x=PronounceGIF, fill = Professional)) +
2. geom\_bar(col = "black", alpha = 0.7) +
3. labs(title="Jiff, Gif, or G - I - F?", subtitle = "The controversial question, answered by Stack Overflow users", caption = "(Based on data from the 2017 Stack Overflow User Survey - Jerry Cheng)") +
4. theme(legend.position = c(.2, .8))
5. theme(legend.background = element\_rect(fill = "white", size = 0.5, linetype = "solid", color = "black")) +
6. theme(axis.line = element\_line(size = 3, color = "grey80"))

Line 1 tells the computer that ggplot needs to be used and a graph should be made from the data frame called “survey\_results\_public”. The x values should come from the PronounceGIF data points, and the filled graph should be “Professional”.

Line 2 creates a bar graph with a black outline color and a lighter opacity.

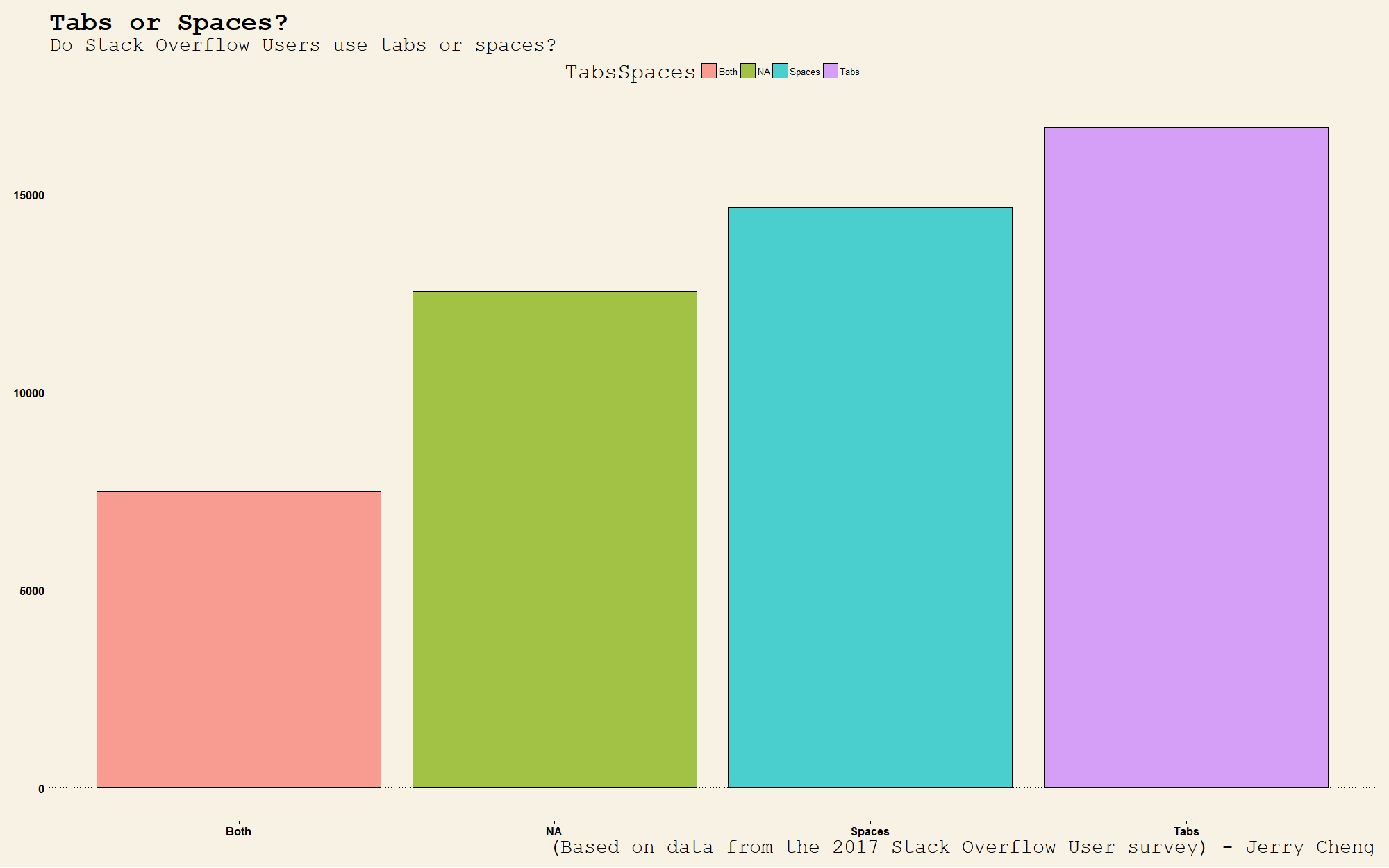
Line 3 customizes numerous labels, including title, subtitle, and caption

Line 4 changes the position of the legend.

Line 5 customizes the fill color, line type, line color, and size of the legend.

Line 6 adds grey lines to each of the axes.

**Graph 5:** Tabs? Or spaces?



Block: ggplot(data = survey\_results\_public, aes(x=TabsSpaces, fill = TabsSpaces)) + geom\_bar(col = "black", alpha = 0.7) + labs(title="Tabs or Spaces?", subtitle="Do Stack Overflow Users use tabs or spaces?", caption = "(Based on data from the 2017 Stack Overflow User survey) - Jerry Cheng") + labs(x = "Tabs or Spaces?", y = "User Count") + theme\_wsj() + scale\_colour\_wsj(“colors6”, “”)

Graph Analysis:

In many programming languages, spacing takes a prevalent role in the usage, functionality, and the legibility of the code. Languages such as Python rely on spacing for functionality, while languages such as Java use spacing for legibility. When it comes to spacing code, programmers are often split between two choices: using the either space bar or the tab key. The results from the 2017 Stack Overflow survey have shown that the userbase is almost split evenly between those who use tabs or spaces, with some even using both.

Code Analysis:

1. ggplot(data = survey\_results\_public, aes(x=TabsSpaces, fill = TabsSpaces)) +
2. geom\_bar(col = "black", alpha = 0.7) +
3. labs(title="Tabs or Spaces?", subtitle="Do Stack Overflow Users use tabs or spaces?", caption = "(Based on data from the 2017 Stack Overflow User survey) - Jerry Cheng") +
4. labs(x = "Tabs or Spaces?", y = "User Count") +
5. theme\_wsj() + scale\_colour\_wsj(“colors6”, “”)

Line 1 tells the computer that ggplot needs to be used, and a graph will be created from data originating from the survey\_results\_public spreadsheet, the x values should come from the TabsSpaces data, and the fill should also come from the TabsSpaces data.

Line 2 tells the computer that a bar graph must be made with a black outline and an opacity of 0.7.

Line 3 customizes labels, including that of the title, subtitle, and caption.

Line 4 customizes the labels on the axes.

Line 5, while technically two different lines, were combined to make one line. This line calls upon the ggthemes library to change the overall theme and color palette of the graph.