STAT W4701: Exploratory Data Analysis & Visualization
Project #1

Overview of data

The data presented for analysis is collected from a survey fielded to students interested in STAT 4701 on January 19th 2016. The goal of the survey was to understand the skills and experience of the class. A summary of the variables covered include:

Variable covered	Data type
Gender	Multiple choice with free text option
Academic program	Multiple choice with free text option
Code/text editor type	Multiple choice with free text option
Experience with tools	Multiple choice
Programming and analytical experiences	5 point scale

Data processing

Initial review and spot checking of the data revealed some inconsistency in the free text typed by survey takers. In order to analyze the data in a consistent manner, free text data were standardized using a script in R.

For example, in *academic program*, entries such as "Applied Math" were classified into the "other masters" category. Ph.D. candidates were classified into one, and responses like "Data Science Certification", "MSDS" and "Data Science" were re-categorized into "IDSE (master)". This was to reduce the number of categories. "Masters" and "Certificate" students were combined since the skills and experiences of the candidates is expected to be similar.

Similarly, *code/text editor type* required re-categorization. "Sublime Text" and "Sublime" were consolidated into one group. "Python", "Ipy" and "Ipython" and "Wrangler" and " Text Wrangler" were also cast into two groups respectively. "Vi/Vm", "NotePad" and "RStudio" were maintained as separate groups. The rest were collapsed into the "other" category since the count for each individual entry was low.

Based on the data structure of the file, we inferred that *experience with tools* can be grouped into the categories to highlight skills. The categorization is shown below:

Skills	Tools indicating experience
R graphics (base; lattice; ggplot2; grid)	R
Reproducible research (sweave; knitr; ipnb)	Knitr
Python	Python
Version control (Git; mercurial; subversion)	GitHub
Databases	SQL
Web frontend	Html; CSS; basic JS; jquery

Data was also collected for other tools, e.g. SPSS, shell, Stata, LaTex, Google Drive and Dropbox. These were excluded from the classification, assuming that tools do not directly coincide with the skills relevant for STAT 4701.

Overview of analysis & findings

After the data was prepared for analysis, plots were created to understand trends in the data.

The demographics of the class were analyzed first. See **Figure 1**. An overwhelming majority of students (79) belong to the Data Science program, of which 60 are male. There are about 17 Statistics masters students, approximately half of whom are female displaying a more equal distribution. 11 students are pursuing other masters programs, while 3 are in QMSS and Ph.D. programs each.

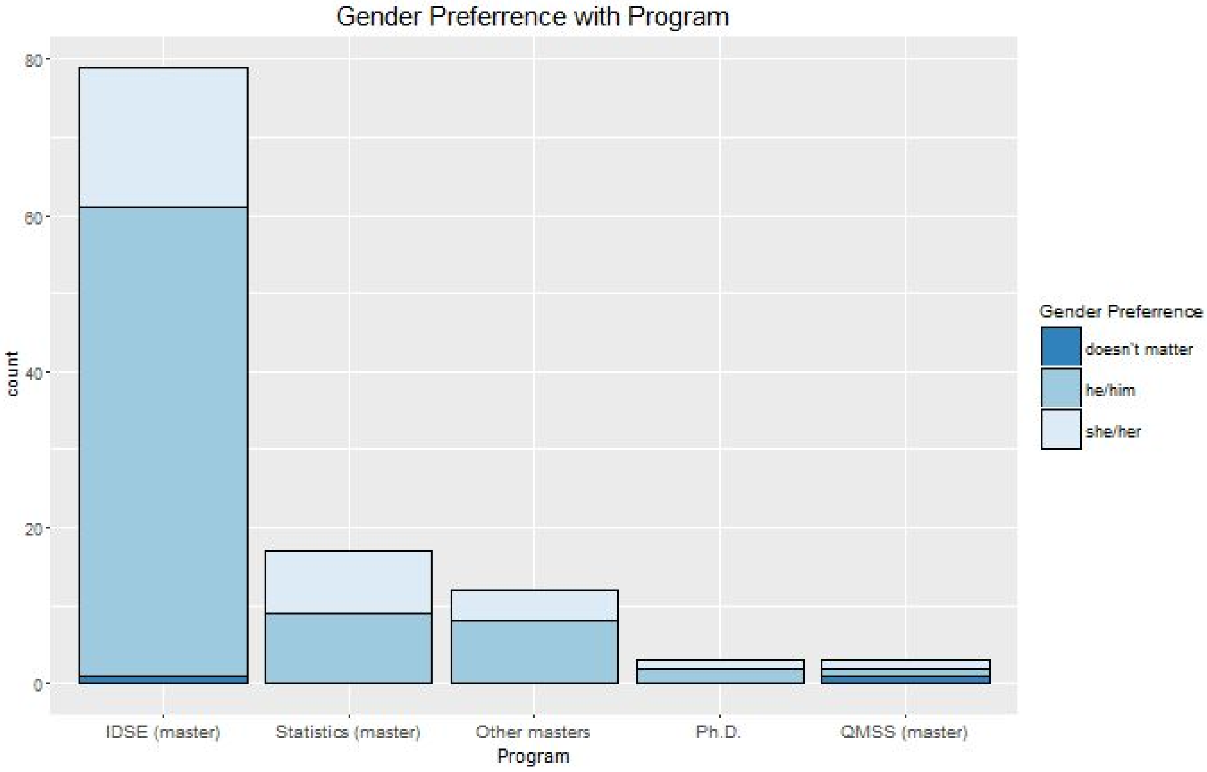

Figure 1*: Demographics of STAT 4701 class*

At the time the survey was administered, there were a number of respondents who were on the waiting list. This is analyzed in **Figure 2** where the number of students confirmed v.s. on the waiting list by *academic program*. it is clear that all the Data Science students are enrolled in the class. This result makes sense especially because this class is geared towards Data Science. There are only two Statistics respondents who are confirmed in the class, while the rest are on the waiting list. Two students from other Masters are on the waiting list while one Ph.D. student is one the waiting list as well.

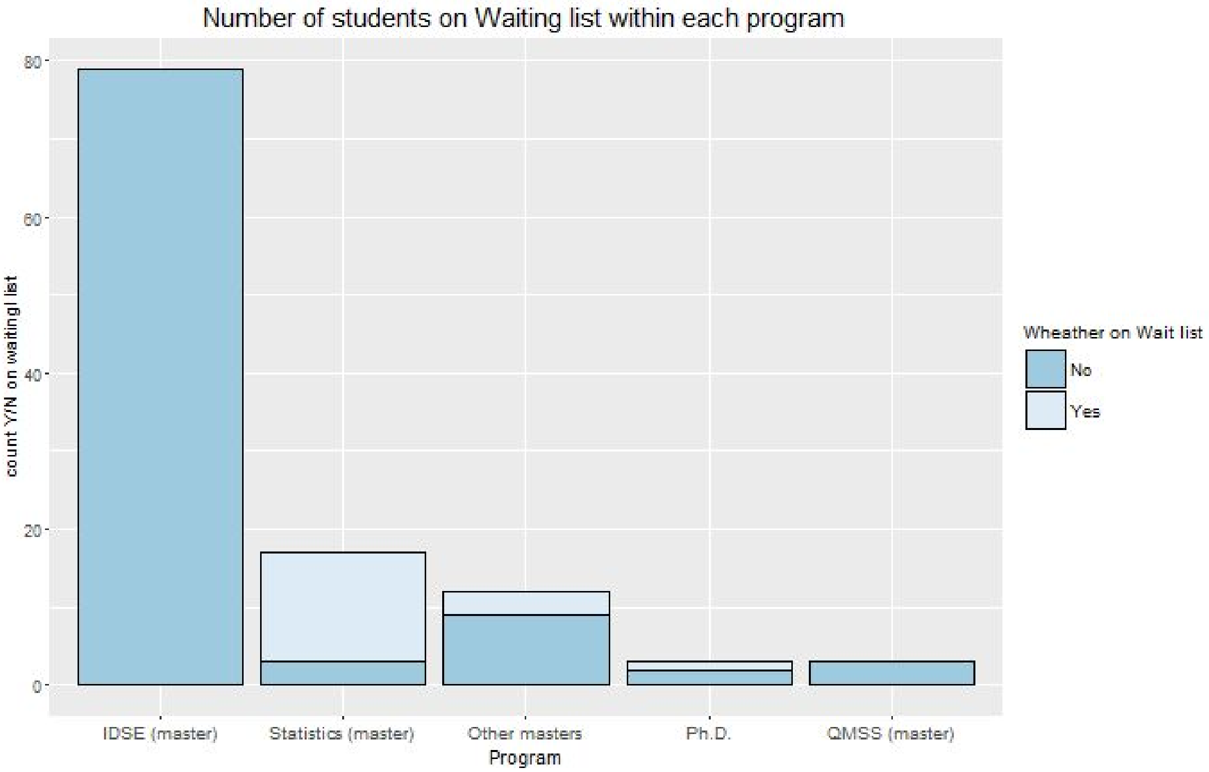

Figure 2: *Number of respondents in each academic program confirmed or on the waiting list*

Figure 3 is a heat map showing the programming and analytic experiences of students and level of expertise. This chart was chosen to provide a quick visual summary of experiences. The conclusion drawn is that R (data manipulation & modeling) is not only the most popular skill, but also the one with the highest number of respondents who are 'confident' or have an 'expert' level. There are very few who have indicated that they have no experience with R (data manipulation & modeling). We can also appreciate that, in general, if someone is confident with R, we can expect that person to be confident as well in most other skills.

By contrast, Matlab column is quite sparse and has only one respondent who is an expert. Majority of respondents are 'a little confident' or have no experience at all.

This chart is also helpful in summarizing the skills of individual respondents. 'Student 72' is an expert at R, R Graphics and R Markdown; and confident in R Advanced, Matlab & GitHub. 'Student 4' and 'Student 9' on the other hand does not have any of these experiences. Another interesting observation is that most students who are confident in R, do not possess more sophisticated skills e.g. R Advanced or R Markdown.

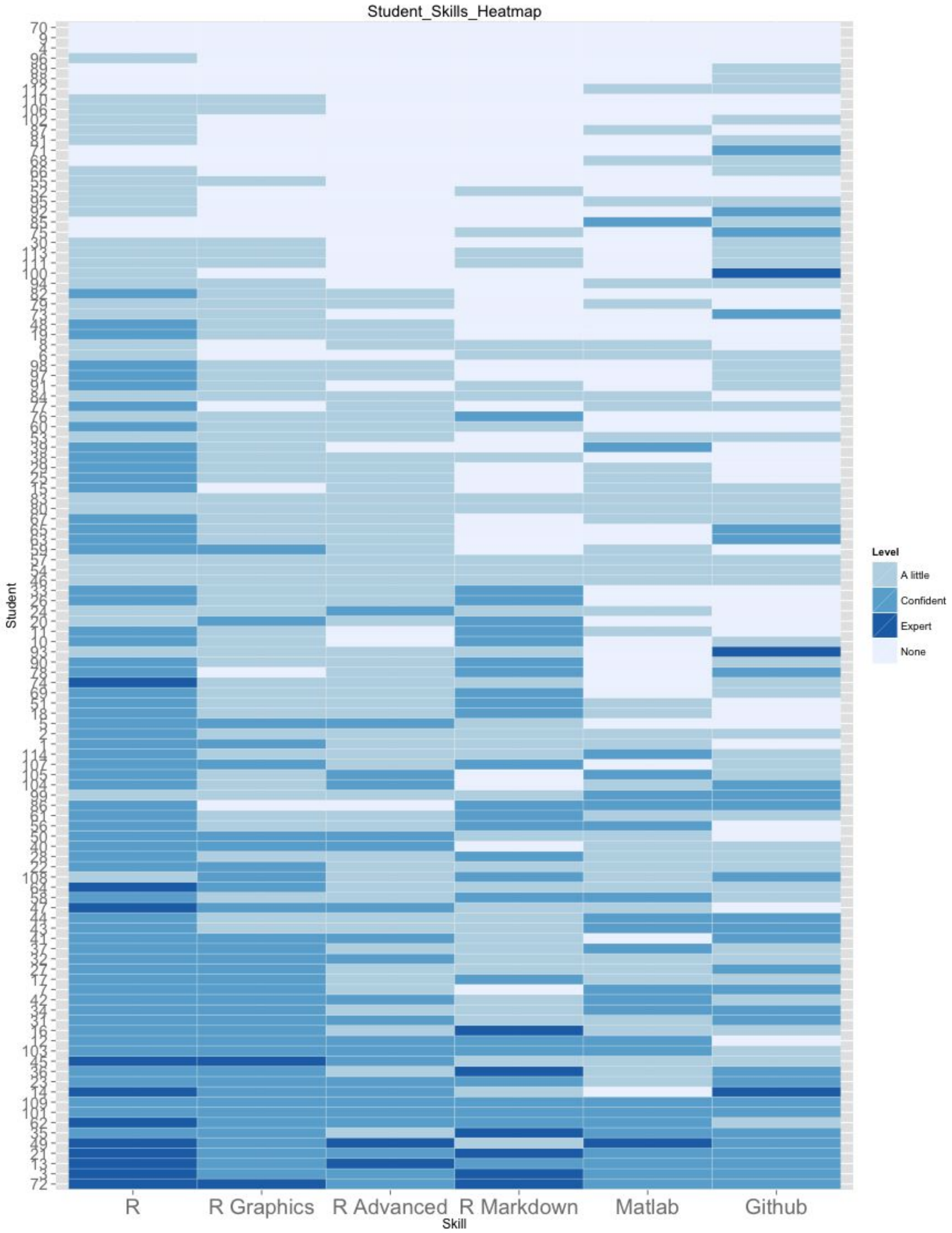

***Figure 3**: Heat map of 'Programming & Analytic Experiences" and expertise level of students.*[1]

[1] **R:** data manipulation & modeling in R. **R Graphics:** Base, lattice, grid etc. **R Advanced:** multivariate data analysis e.g. spatiotemporal data, visualization & modeling. **R Markdown:** reproducible documentation **Matlab:** data manipulation, analysis, visualization and modeling. **GitHub**

Figure 4 suggests that most of the respondents in the class are familiar with R Graphics and Python. Less than half the respondents are familiar with Databases and Version Control. R Research and Web Frontend are skills possessed by only a handful of students.

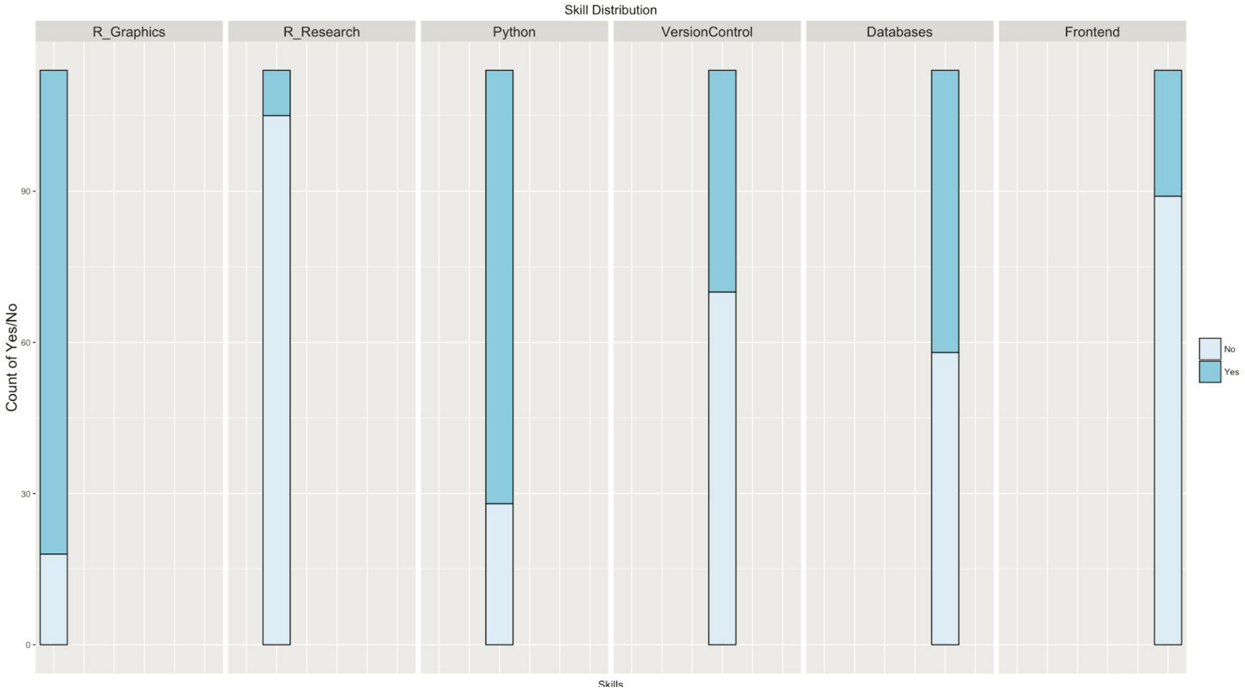

Figure 4: *Software experience of respondents*

The number of students in each *academic program* by *software experience* is shown in **Figure 5.** It can be seen from the figure that the number of students with skills varies considerably by academic program. Respondents in all five categories are familiar with R Graphics. Most students in Data Science and all students in QMSS program are acquainted with Python. Version control is a skill some students in all groups possess, though a greater number of students in Data Science and Statistics are familiar with it. A small number of students know web front end. A number of students are familiar with databases, though most of them are in the Data Science program.

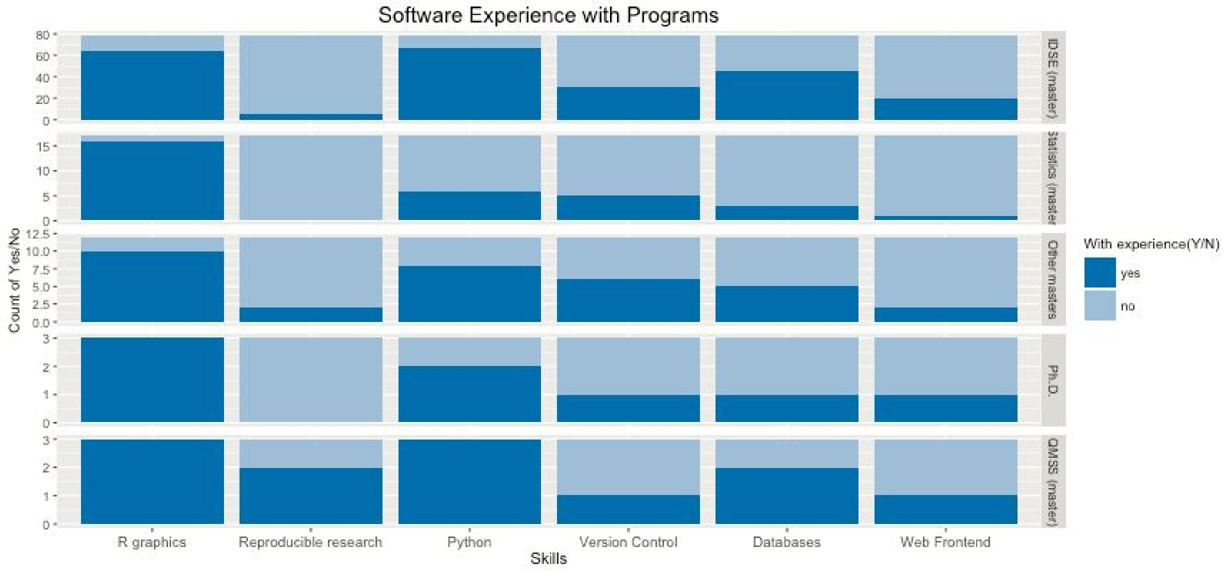

***Figure 5**: Count of students with software experiences grouped by academic program*

Figure 6 is a gender wise distribution of skills. Percentage of males who are familiar with Python, Version Control and Databases is greater than percentage of females while the trend reverses for R graphics packages.

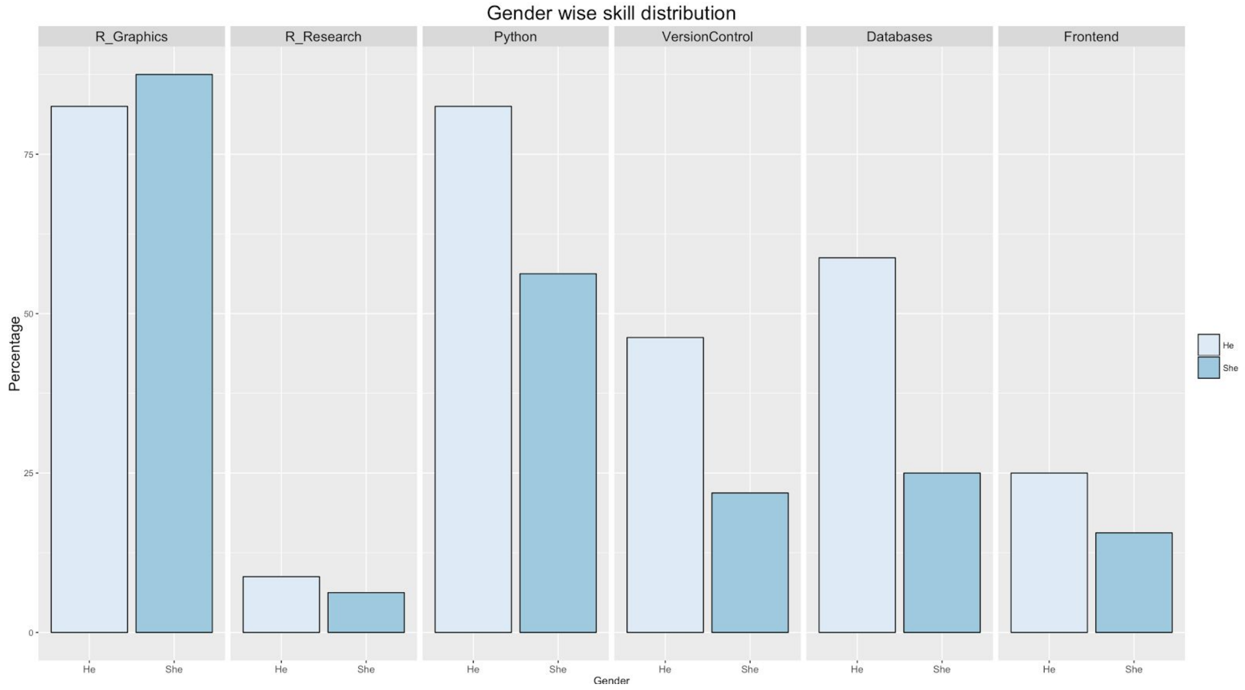

***Figure 6:** Count of students with software experiences grouped by gender*

The bubble chart in **Figure 7** helps to emphasize the difference in size of each skill group, per program. We can see that, by far, the greatest groups are those comprised by IDSE master students who know R, R graphics and Python, respectively. On an intra-program level, the same conclusion can be derived for each program, except the Statistics master, where students are more confident with Matlab than with Python. Finally, it can also be observed that even the smallest skill group of IDSE master students (web frontend) comprises more people than any other skill group from any other program.

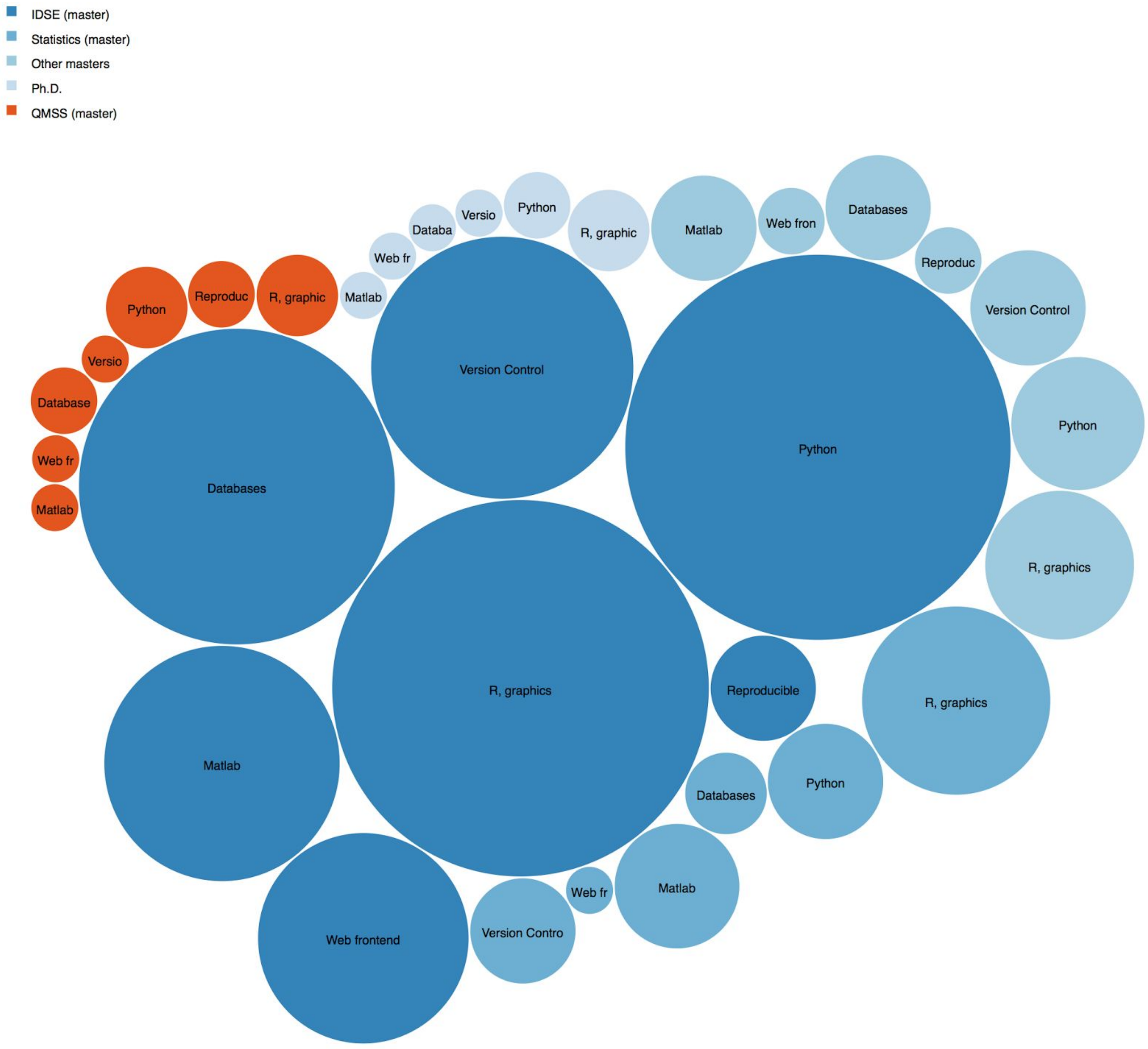

__Figure 7__: Bubble chart of number of respondents with various programming experiences by academic program

The same information can be looked at by proportion in **Figure 8**. It gives a clearer picture that all of QMSS and Ph.D. candidates have experience with R Graphics. QMSS students also display the highest proportion of students with Python experience. Statistics and Other master candidates display the greatest proportion of students with Matlab experience. It is interesting to note that Statistics and Ph.D. students don't have any 'Reproducible tools' experience. This experience is also generally the lowest for each *academic program* compared to other tools' experience.

Since most of the respondents in the class are Data Science students (as seen in **Figure 1**), it is helpful to take a closer look at the skillset of that group. Greater than 75% of the Data Science respondents are familiar with Python or 'R Graphics'. More than 50% are familiar with SQL. Around 40% are familiar with 'Version Control' tools. A very small proportion of Data Science students in the class are familiar with 'reproducible tools'.

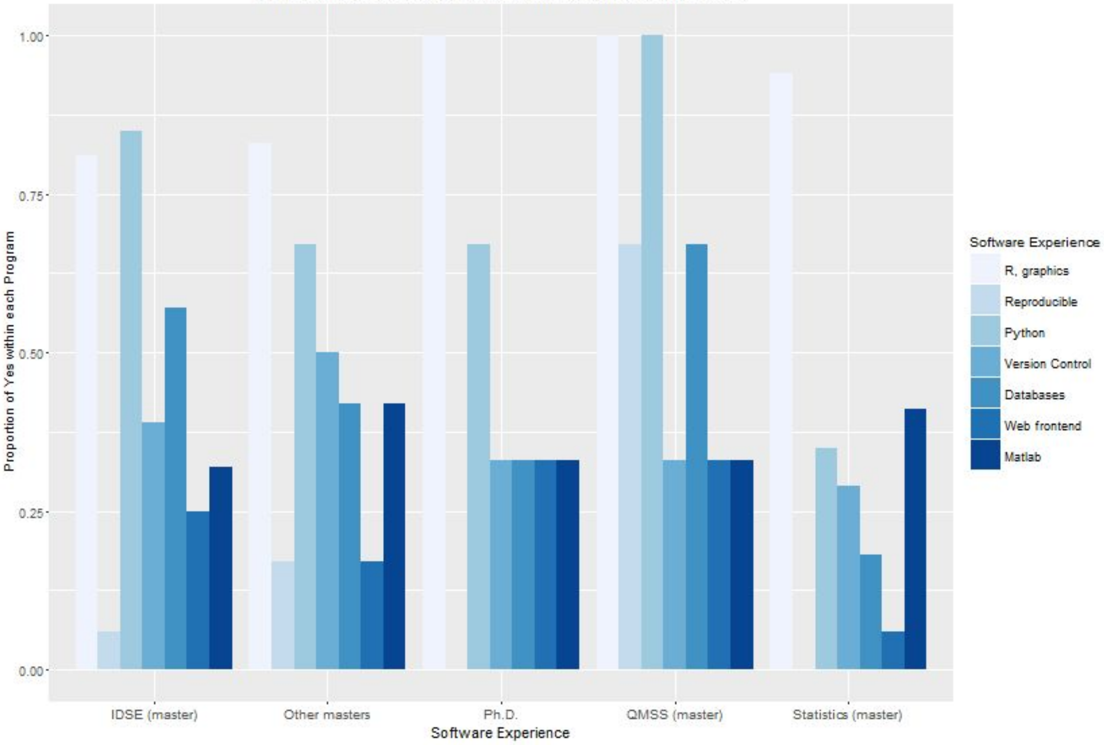

***Figure 8**: Percentage of students in each academic program with software experience*

The software experience of respondents based on if they are enrolled in the class or on the waiting list would be an interesting cut of the data. This is shown in **Figure 9**. It seems like most of the respondents there are a number of students on the waiting list who can bring Matlab skills to the class. However, proportionally, respondents on the waiting list lack experience with Python, Databases and Web front end. This result makes sense especially because it is clear from **Figure 2** that most of the students on the waiting list are in the Statistics program.

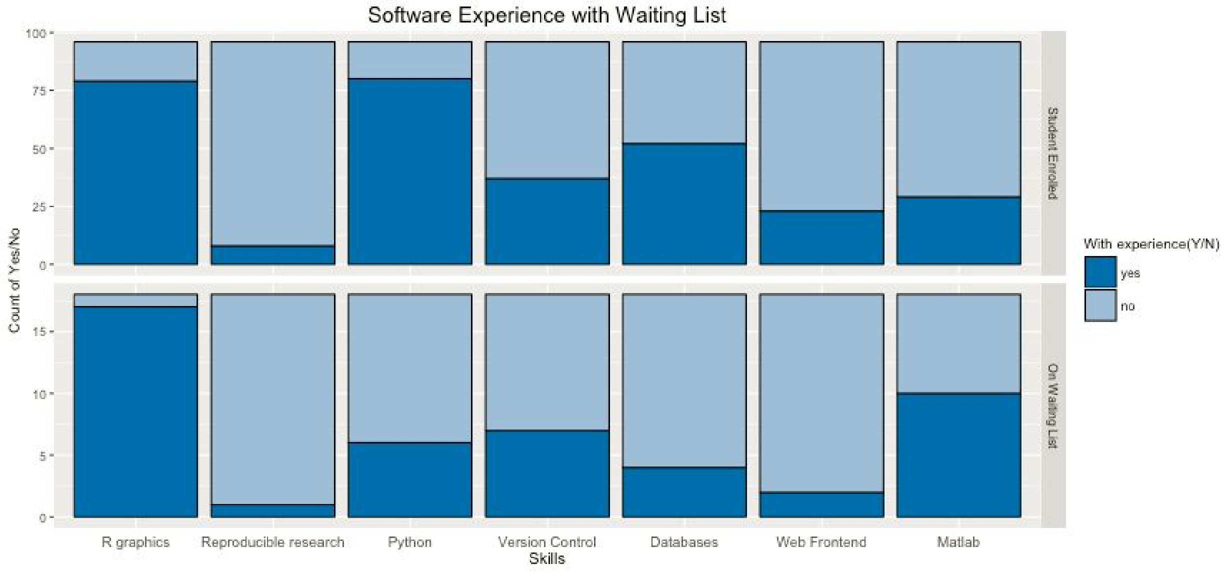

***Figure 9**: Number of respondents with software experience by confirmed or waiting list*

It is clear from **Figure 10** that RStudio is the most popular text editor in the class. More than 62 respondents are familiar with it. Sublime Text is the second most popular editor with 17 respondents who are familiar with it. Vi/vim is used by only 11 respondents, while 10 or fewer respondents have used other text editors.

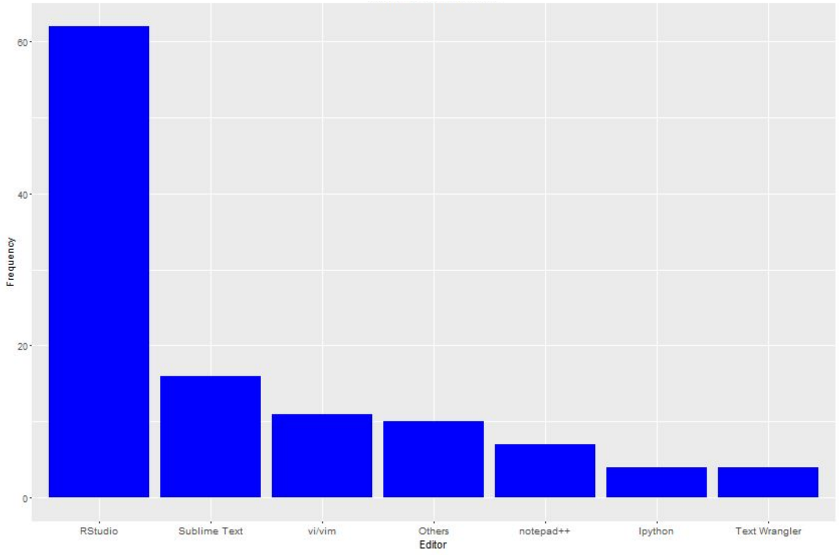

***Figure 10**: Frequency of text-editor use*

Figure 11 shows the preference of text editor use by program. It is clear that RStudio is preferred by respondents in all programs. RStudio is strongly preferred by Statistics students. Half of Data Science students use RStudio, while the others use a variety of editors including Sublime Text, vi/vim, notepad++ , Ipython and Text Wrangler.

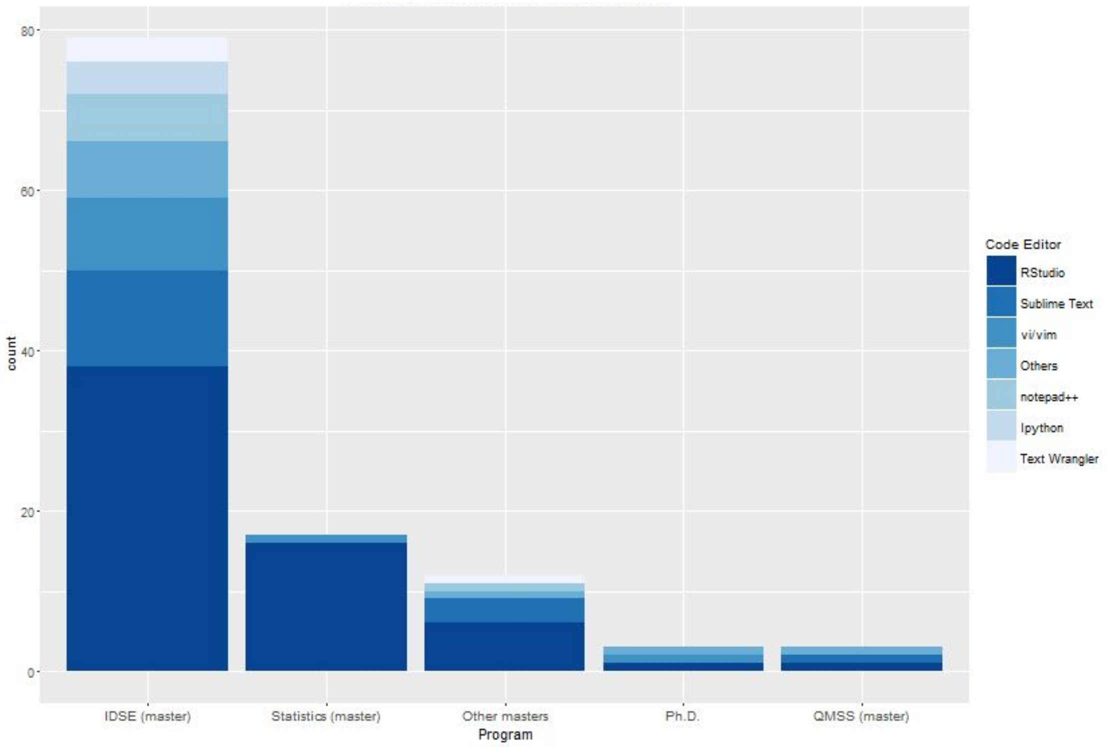

Figure 11*: Text editor preference by academic program*

Finally, two more additional, interactive plots have been done using D3.js. These plots provide the same information as the plots in **Figure 5** and **Figure 8** in a much more interactive way, and can be found in the *html_code* folder.