HW1 Visualization. The Class Survey Data Visualization.

The Internet Explorers

Note

We prefer this report is read in its html version for a better interaction with the reader. We can find the file HW1 Report.html in the directory provided.

Introduction

This study analyzes the data from the initial survey conducted at the beginning of the course "Exploratory Data Analysis and Visualization" at Columbia University. The purpose of the first project, as the team members agreed, is to do some hands-on work with the visualization tools introduced in lecture. We will present, with original visualizations, answers to some questions about the data. To draw some of the charts, we copied the format from fivethirtyeight.com, which is a great data-analysis and visualization site that we took as inspiration. Also, the colors used in the charts are the same as those used in the Google docs. This could be easily implemented using the package ggthems as a complement of ggplot2. We strongly recomend that the reader take a look at these to avoid manually changing the font, the sizes, the colors, etc... in each graph individually in your future presentations.

Who we are

In order to have a better understanding of the makeup of survey responders, an interactive visual graph was developed using d3 to show the the gender distribution among our various programs. Clearly, the number of male students (71%) is more than female (28%).

Moreover, male students dominate the population for each program. It is worth mentioning that the highest percentage of female students is within the Masters in Statistics program, with around 47%. On the other hand, female students have the lowest percentage in the Data Science Program (both for Masters and Certification). Overall, the percentages can be summarized in a symmetric graph as follows. An interesting interactive chart can be found here.

How old are we?

We thought that it would be interesting to have the age of the students in the course to understand more deeply their demographic information. Hence, we downloaded available pictures in courseworks and sent them to Face API from Microsoft Project Oxford to estimate the age of each picture. By doing that, we have a relatively accurate age estimation of 31 students, wich is more than 28% of the entire population. For those interested in how we did it, aside from the code in Github, there is this useful tutorial: Analyzing 'Twitter faces' in R with Microsoft Project Oxford. This visualization is intended only to be reviewed within the activities in the course. If you would prefer your picture not to be there, just provide us that feedback and we will remove it immediately. Also, we would like to know if the estimation is accurate or not. So, if you don't mind letting us know, please enter your true age in the following: form. In the html version, the reader can slide the mouse over the pictures to see the estimated age.

The histogram of frequencies, shown in Figure 2 in next page, indicates that the male students seem to be, on average, approximately 5 years older than the females. The mean age values by gender, 27.6 for females and 32.7 for males, is traced with a vertical dotted line.

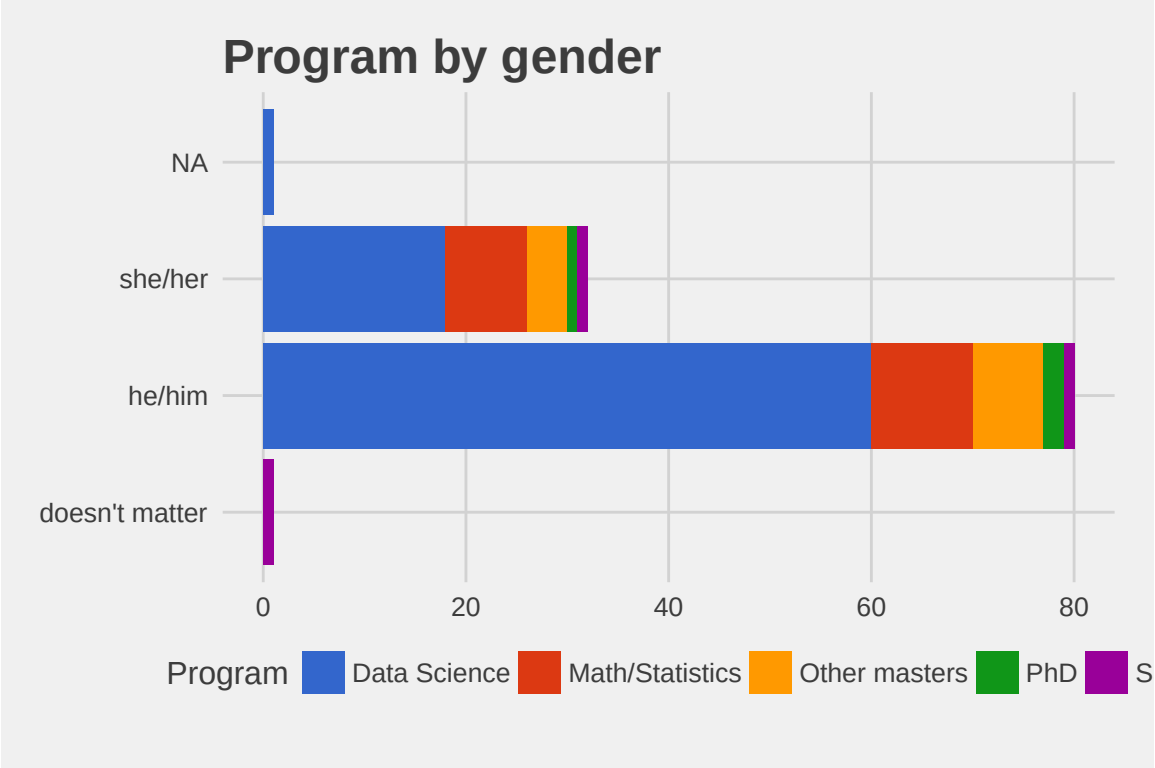


Figure 1: Student programs by gender

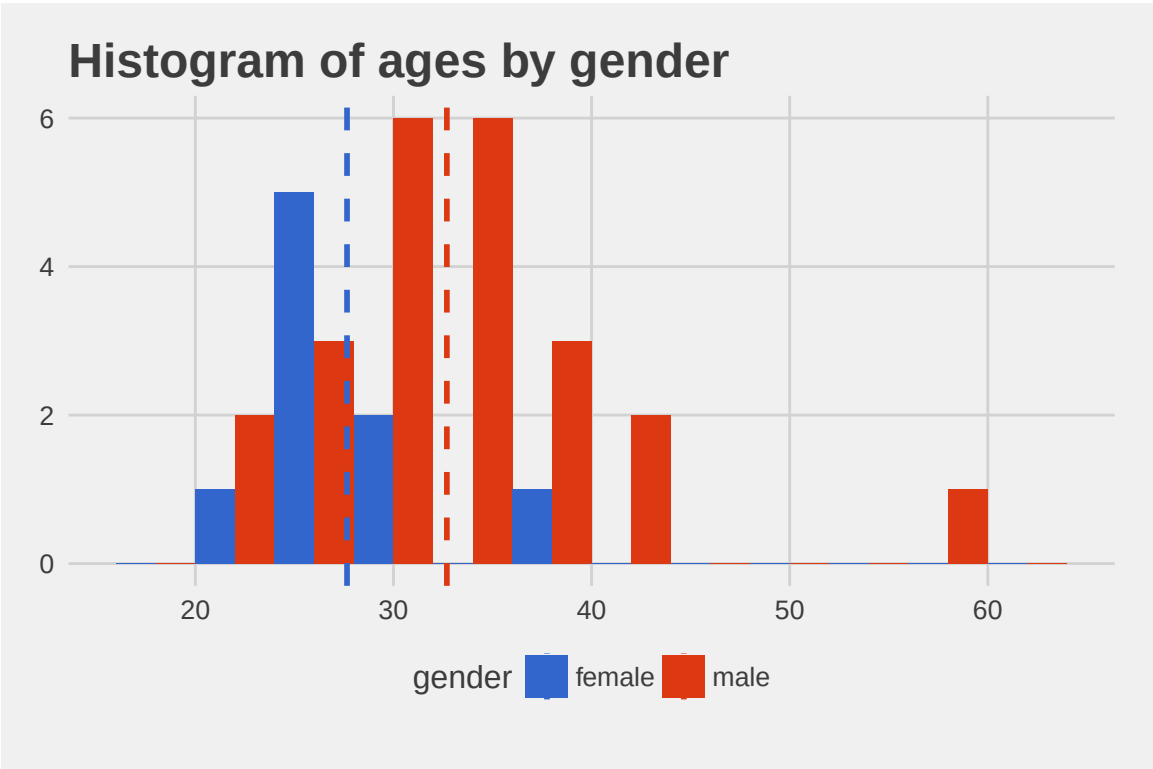


Figure 2: Age distribution by gender

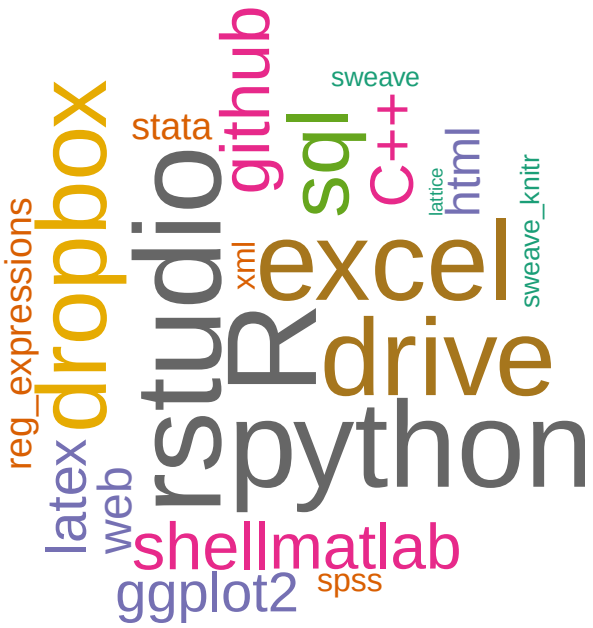


Figure 3: Wordcloud of most mentioned tools

Familiarity with Computational Tools

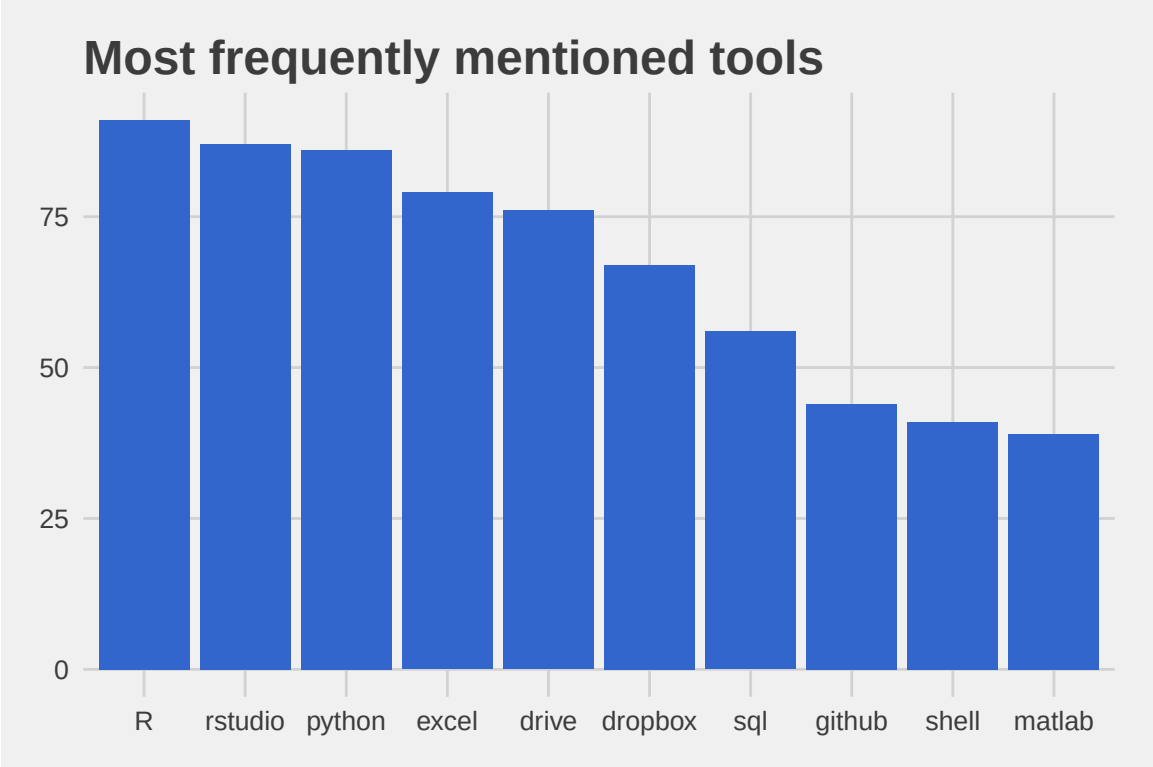
In this section we explore methods to visualize the relative knowledge of the class as far as computational tools are concerned. In the survey, students were asked to select which tools they felt comfortable using from among a list of 20. The type of tools varies from specific computer languages like Python, to more general applications like Google Drive.

We present two methods for visualizing which tools the greatest number of students feel comfortable with. The first is a Word Cloud. The Word Cloud below prints words with a font size that is correlated to the frequency of responses. So, for example, if more students selected 'R' than any other tool, then 'R' would show up in the largest font. To obtain this graph we used the package wordcloud.

The Word Cloud, while not quantitative, provides a nice visual to give the audience a general sense of which tools are most well known among the class. We see that 'R' and 'Python' show up frequently, as well as 'Excel' and 'Drive'. We were not surprised by this result, since these languages and tools are commonly pushed in our classes.

For a more quantitative visualization of the responses, we also present a histogram which is ordered from largest number of responses to smallest. From this we can see more accurately than with the Word Cloud which tools were responded to more frequently than others.

The conclusion one can draw from this visualization of our comfort level with various tools is that the majority of the class has experience with 'R' and 'Python'. Some of the more specialized tools like 'SQL' and 'Github' are less familiar to the class, and certain other tools such as 'Sweave' and 'grep' do not show up at all because few people are experienced with them. These considerations would be important for an instructor wanting to determine what class members know coming in.



What else do we know?

To explore the data about programming expreience and skill we show a bar chart and a series of bubble plots.

These folowing plot show the distribution of programming level among different skill levels.From the bubble chart, we can find that few people consider themselves as an expert in a specific field. More people have none experience in Matlab when compared to other fields.We can also find that people have more experience and are more proficient in modeling when compared to other fields.

From the graph Matlab versus data manipulation and modeling, we can find that people who are good at data manipulation tend to have little experience in the filed of Matlab. From the graph Matlab versus Github, we can find that the level of people in the field of Matlab and Github are quite similar. Most of them have limited knowledge in these two fields. From the graph graphics.basics versus documentation, we can find that the level of people in the field of Matlab and Github are quite similar. If people are confident in one field, they tend to be confident in the other field. If they have limited knowledge in one field, they tend to have limited knowledge in the other field.

This is the end of our presentation. We hope you have taken something useful from it.

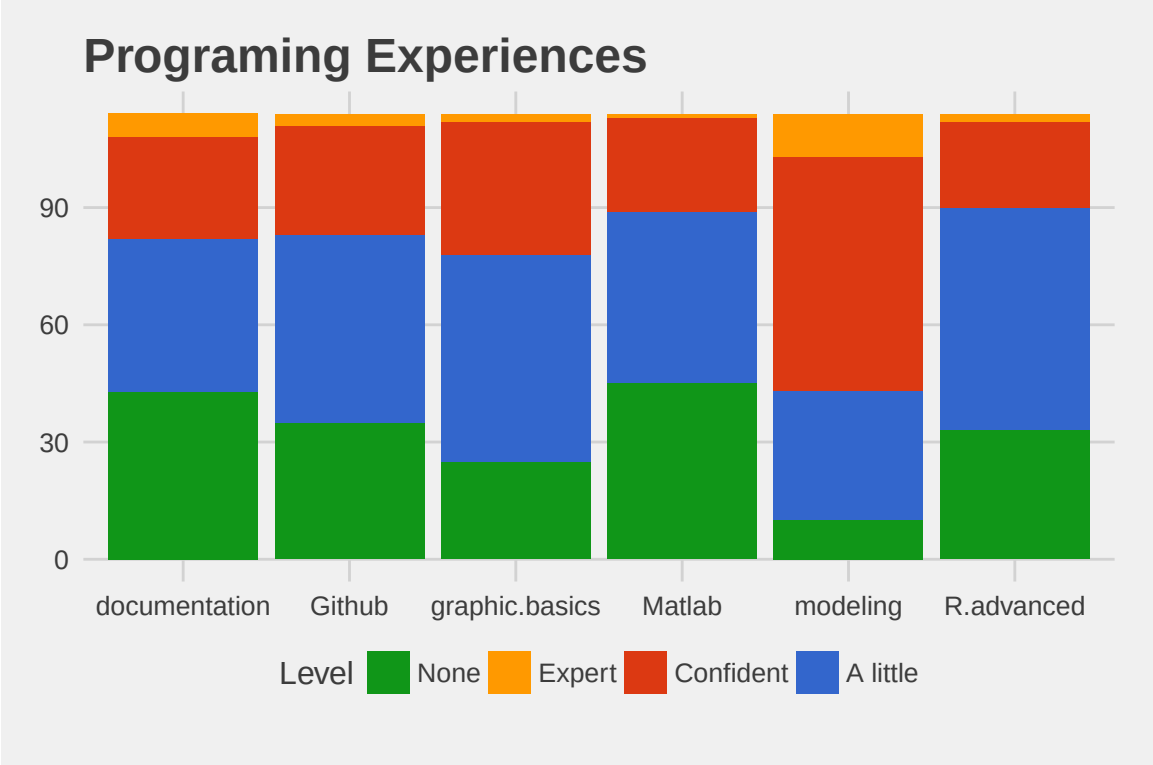


Figure 4: Most mentioned tools

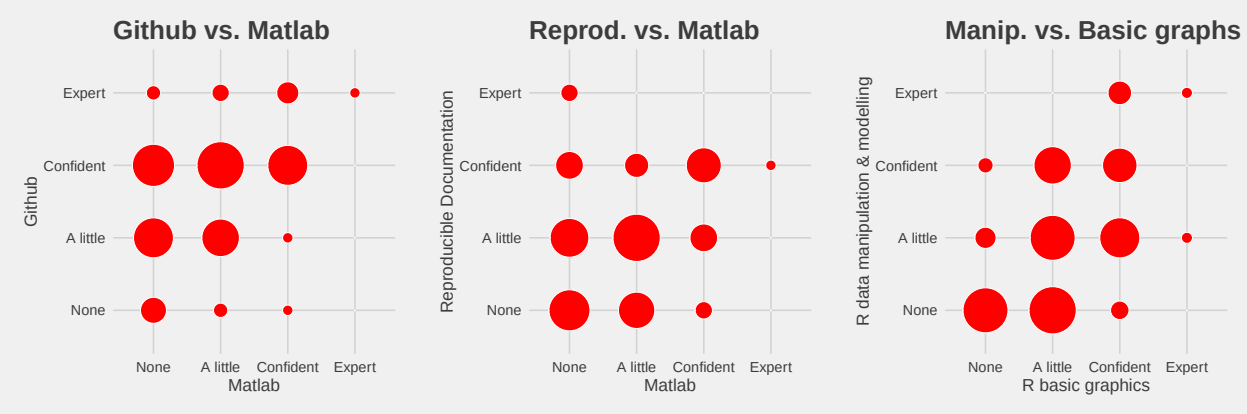


Figure 5: Bubble plot with knowledges and experiences