HW1 Visualization. The Class Survey Data Visualization.

The Internet Explorers

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. Its purpose, as the team members agreed, is to do some hands-on work with the visualization tools introduced. We will present with an original visualization, 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 google docs. This could be easily implemented using the package ggthems as a complement of ggplot2. We strongly recomend that you take a look at them if you don't know them already, to avoid changing manually: 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 makeup among out various programs in terms of percentages. Clearly, the number of male students (71%) is more than female (28%). Moreover, male students are dominating the population for each program. It is worth mentioning that the highest percentage of female is within the Masters in Statistics program, with around 47%. On the other hand, female students has 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 thught that it would be ideal to have the age of the students in the course to understand more deeply their demographic information. Hence, we downloeaded available pictures in courseworks and send 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, besides the code in Github, there is this usefull 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 think that your picture should not be there, just provide a feedback and we will remove it inmediately. Also, we would like to know if the estimation is accurate of not. So, if you have 2 minutes, we will appreciate your feedback filling this form. You can slip the mouse over the pictures to see the estimated age.

The histogram of frequencies, showd in Figure 1 below, indecates that the male students seem to be, in average, aproximately 5 years older than the females. The mean age values by gender -27.6 females and -32.7 years old for males- is traced with a vertical dotted line.

some closing text..

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

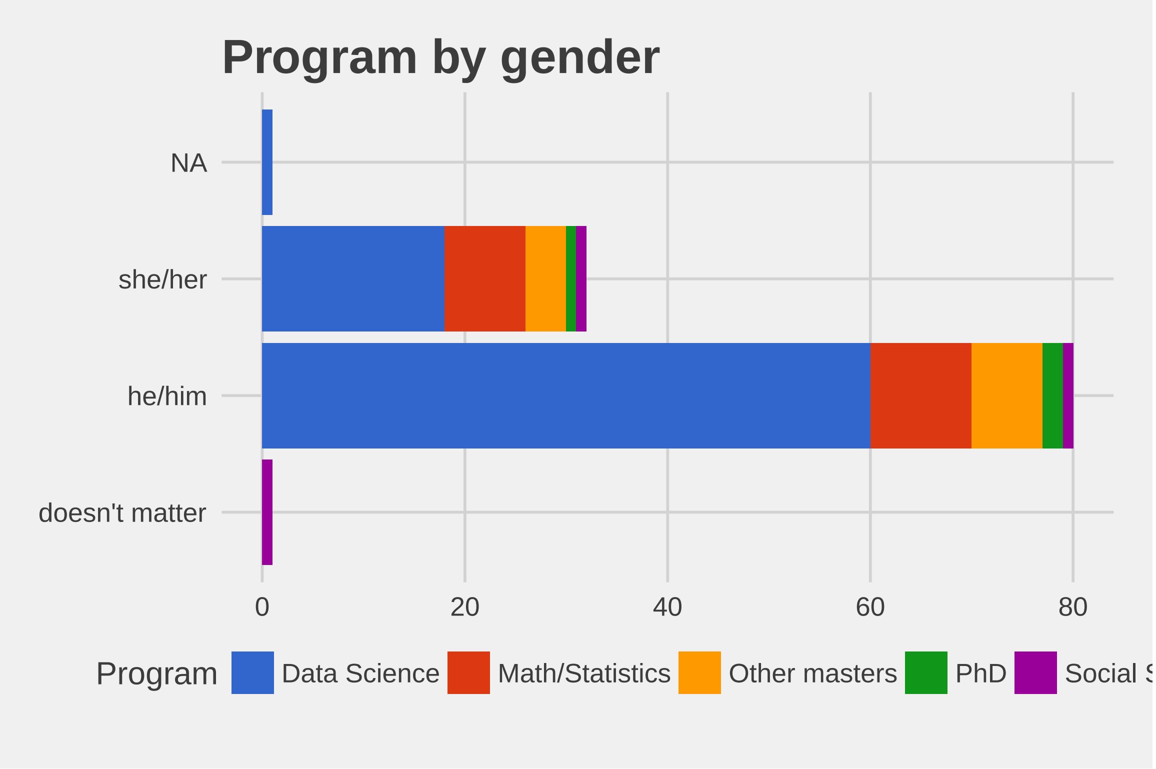


Figure 1: Figure 1

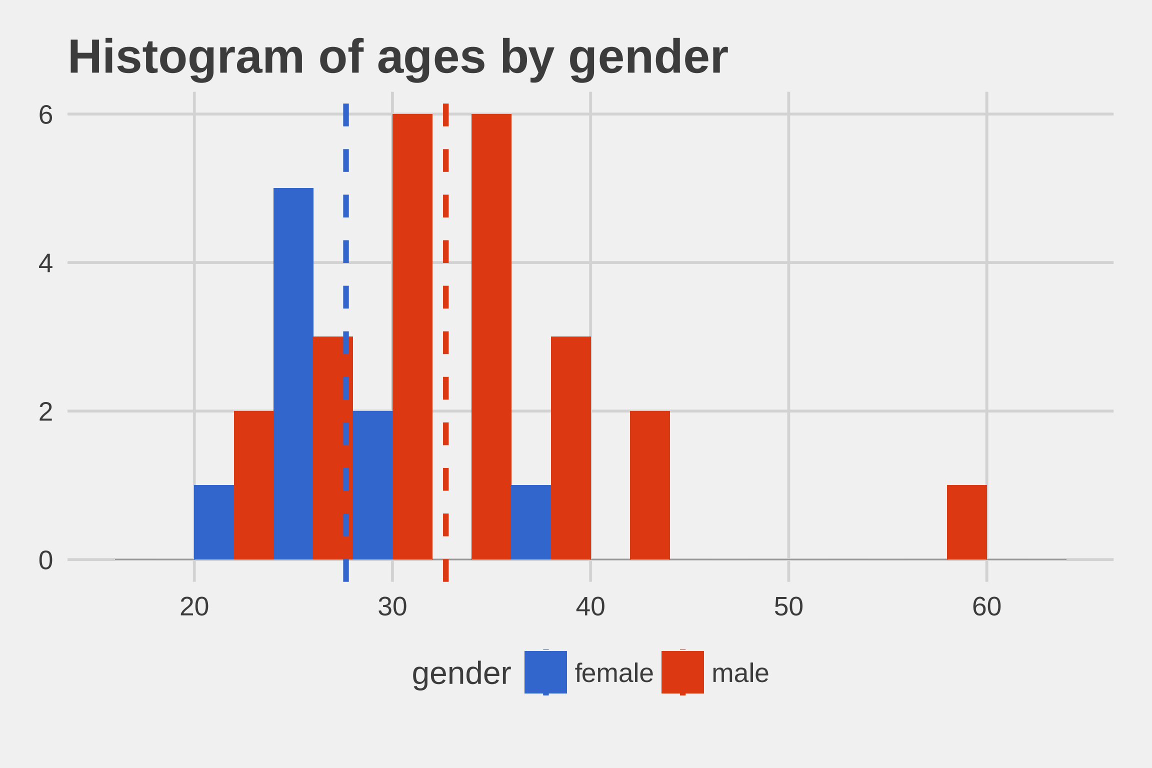
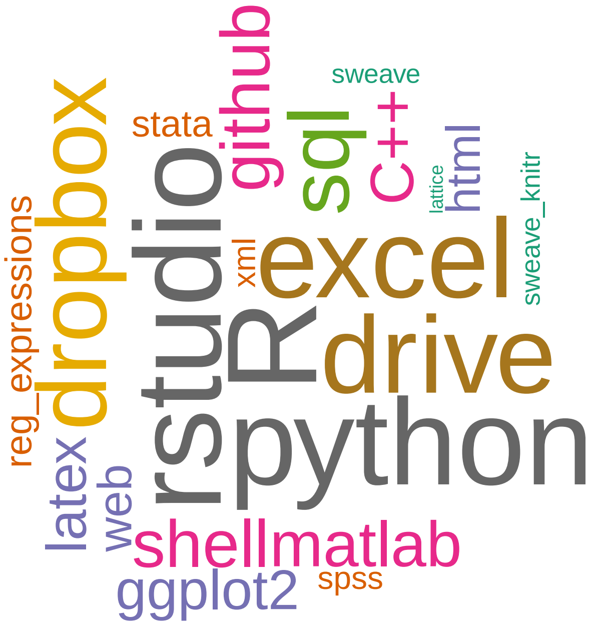


Figure 2: Figure 1

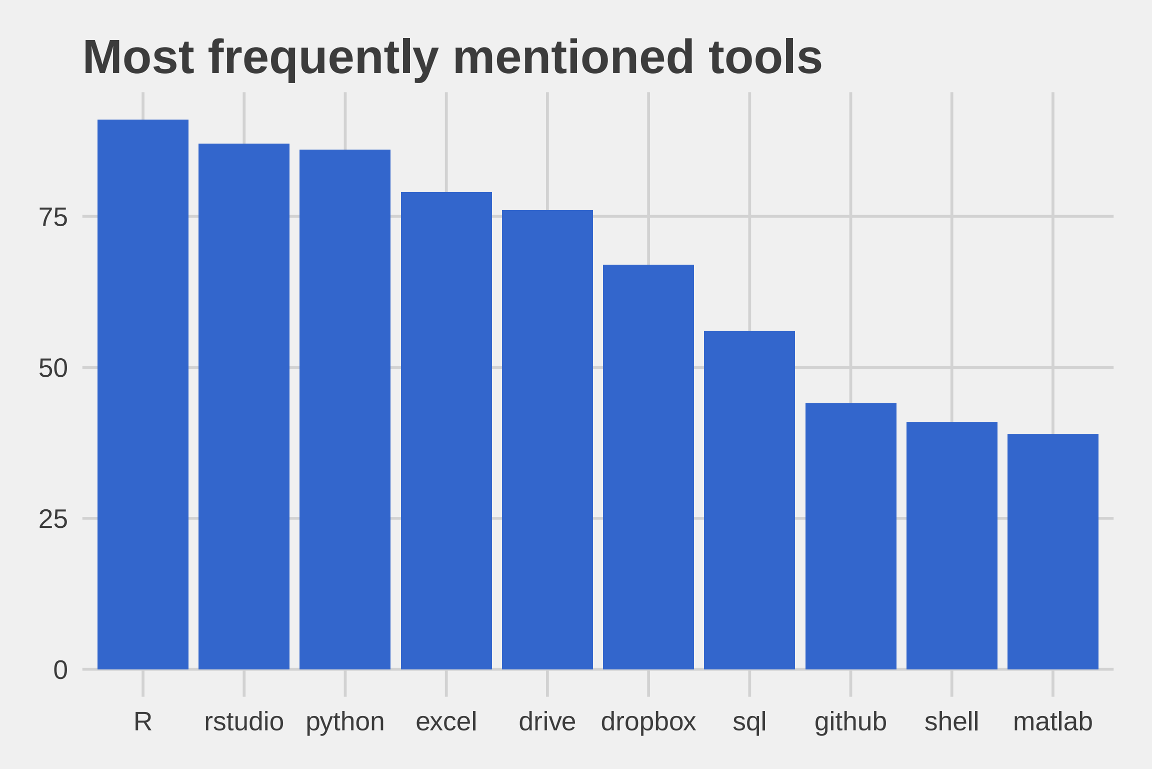
among a list of 20. The type of tools varies from specific computer languages like Python, to more general applications like Google Drive.

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 do we know

To explore the data about the programing expreiences and skills we trace a bar and a series of bubble plots.

This plot shows the distribution of programming level among different skills. 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 the presentation. We hope you take something usefull from it.

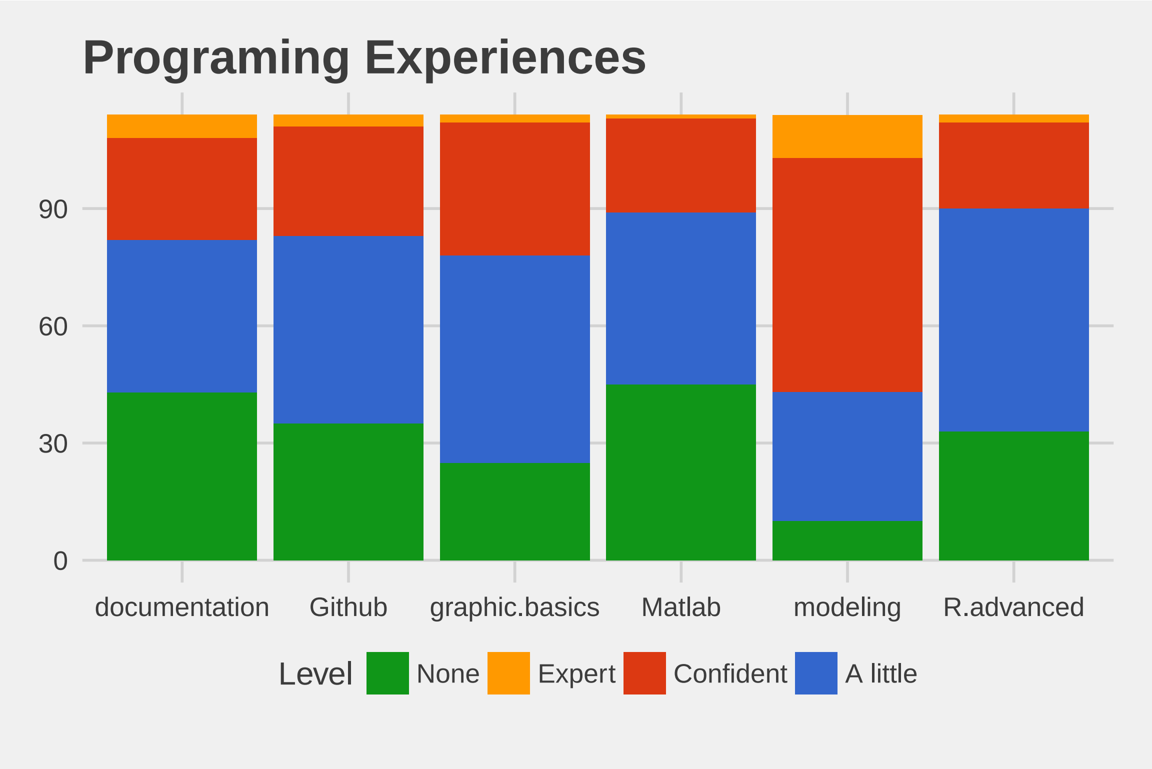


Figure 3: Figure 1

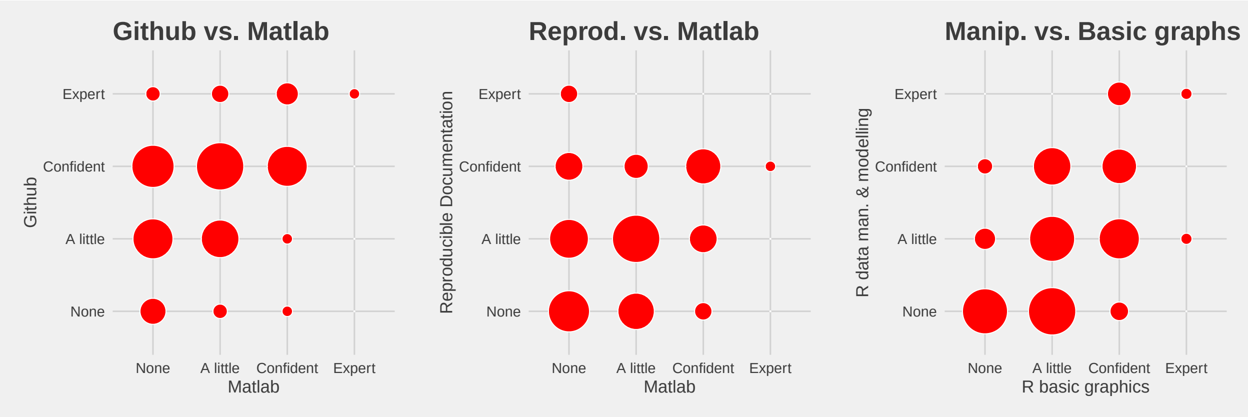


Figure 4: Figure 1