Project 1: The Class

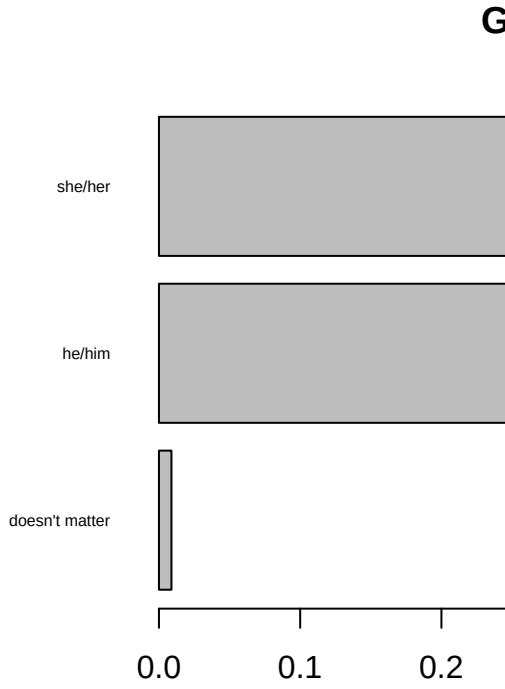
Background

The following analysis examines the 114 survey responses from our EDAV class.

Waitlist

- 16% of the class started on the waitlist.

Gender

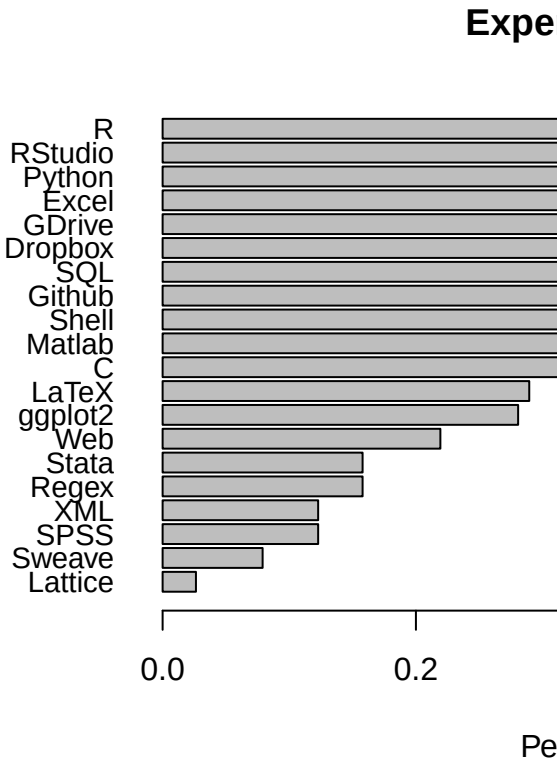


- 71% of the class is male, 28% female, and 1% says their gender doesn't matter.

Program

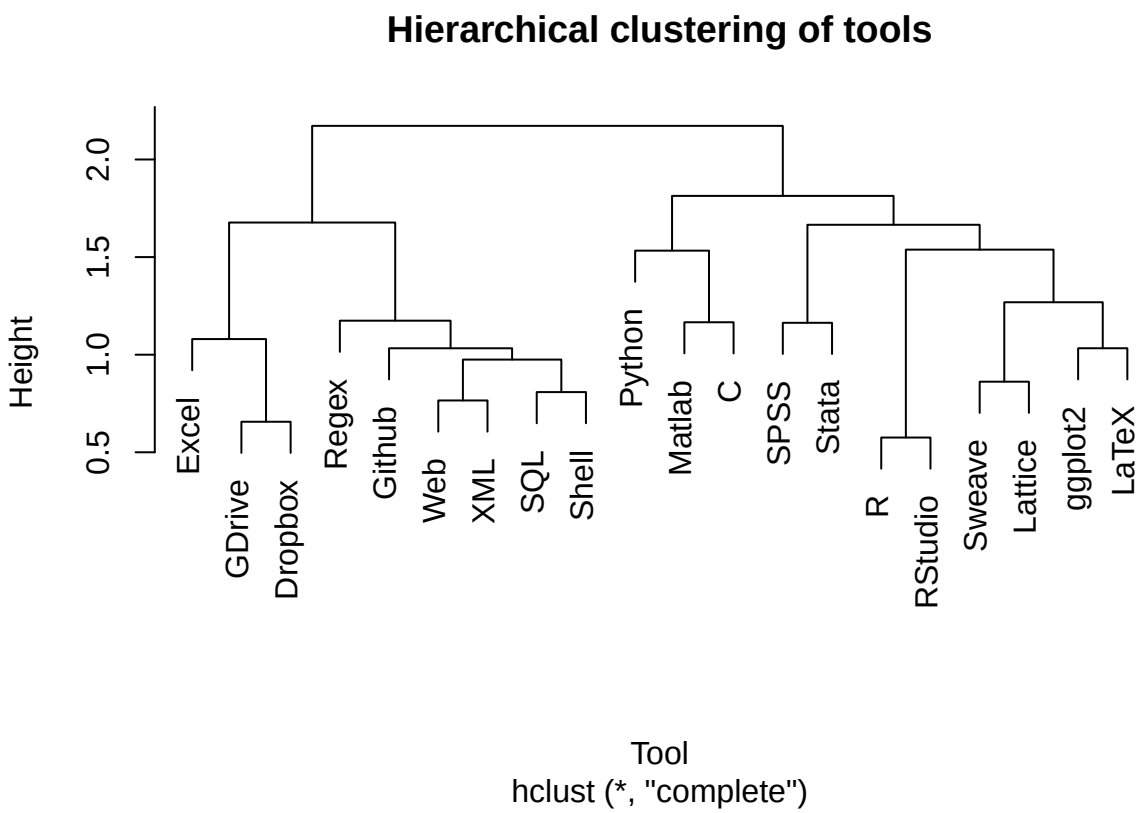
- Looking at the distribution of students by program we see that almost 50% of our students are Master in data-science, and 20% are in certificate program.

Tools



- R, Python, and Excel are the most commonly used tools among students.

Hierarchical clustering reveals distinct clusters of tool usage

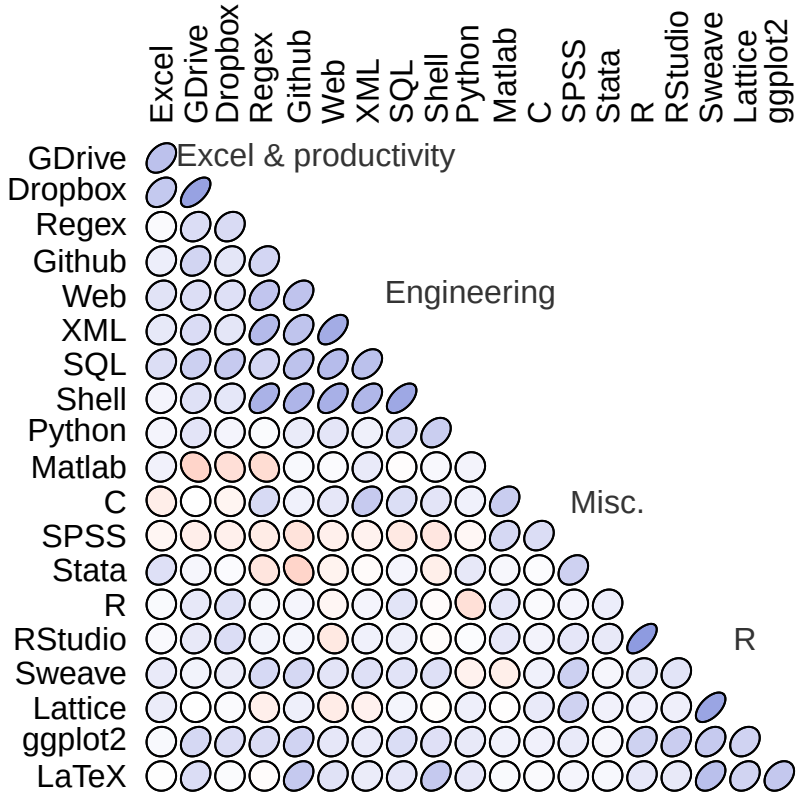


We interpret these clusters as:

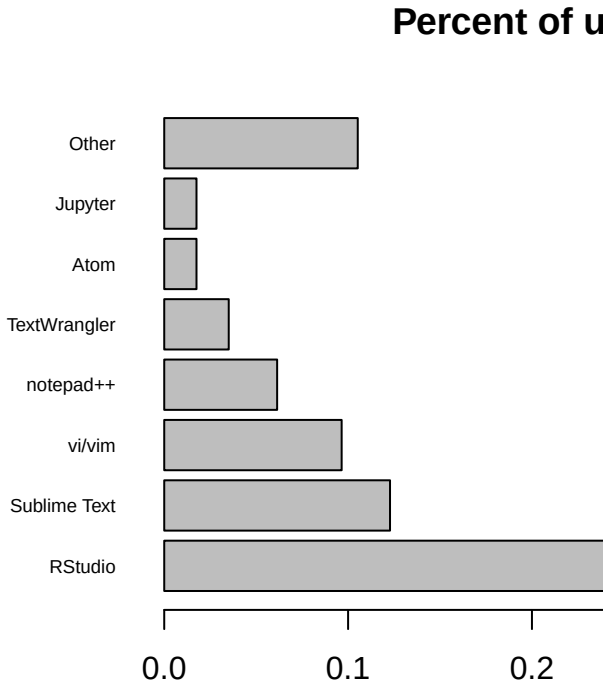
- Excel & productivity tools
- Engineering tools
- Miscellaneous languages
- R and R packages

A correlation plot reveals the granular relationships between tools:

- There are some competitive products that are positively correlated while others are negatively correlated. For instance: dropbox and google drive are positively correlated while R and Python are negatively correlated. Perhaps it's because dropbox and google drive have similar usability while R and Python are less interchangeable.



Text editors



- The distribution of text editor shows that RStudio is heavily favored.

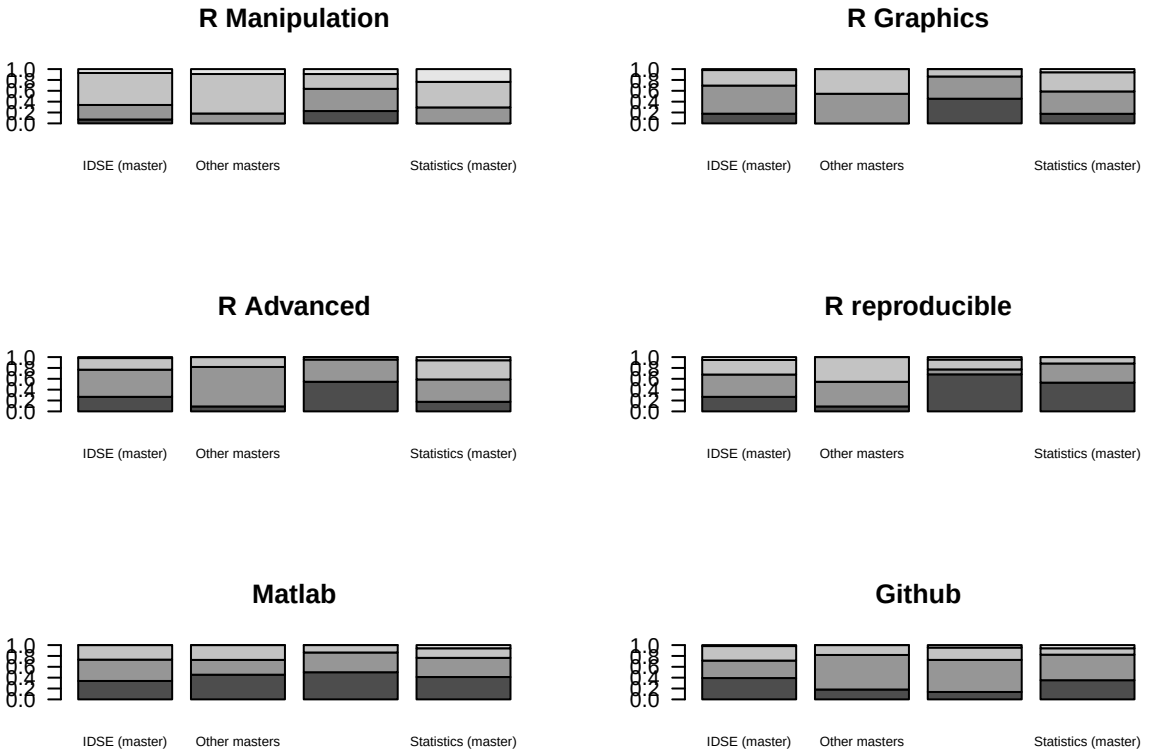
- The sankey chart shows that the preference for RStudio is explained by the large number of statistic and data-science masters students.
- We can see text editor preferences by program: few students of statistics prefer different editors while 40% of data-science students do not use RStudio.

Experience by tool

- When it comes to R, students are more experienced with manipulation than with other skills.
- Students have little to no experience with most of the tools below

Generally, for each skill, there are few people who are experts of that certain skill. Most people (more than 2/3) know nothing and a little about R advanced, R reproducible, Matlab and Github, while more people are confident about R manipulation and R graphics. In general, it's a good mix of all kinds of skill sets.

Tool expertise by program - Generally, Data Science students know less about all these first five tools. But they are pretty good at using Github. - Other masters other than data science are good at R manipulation and R graphics. At least, none of them know nothing about these two tools. - Among all these six tools, experts are minority. Most of them are in statistics major. - Most of students, whatever his/her major is, only know a little about them. It means that we are far from to be a qualified data scientist.



The figure above helps us understand the experience distribution program-wise, categorized by 4 ordered levels (None < A little < Confident < Expert) . A few inferences that can be drawn from the graph are as follows :

1. A majority of class is confident in data manipulation in R with very few individuals in the none category and expert category. But its noteworthy that data manipulation in "Expert" category is the highest value among all the expert categories.

2. A majority of students have “little” to “none” expertise in producing reproducible research and Matlab. Hypothetically the instructor should focus on these areas.

3. There is an almost equal distribution of students who are confident in Github and those who know “none” about it.

This Chart shows a relation between *program* , *tool* and *experience level*. We removed the association between program and tool when the level of experience is None. In Master Data-Science (IDSE) Github and Matlab are the most unknown topic and in certification program have a poor experience with R.reproducible. However 30% of all students do not have any knowledge in Matlab and 91% of them have worked with R.Manipulation. Analysis on Tools: