EDAV Project1 Part by Using Python

Introduction

In this part, we used python to do data extract, load, transform (ELT) and processing. Then we visualized average knowledge of each skill for different groups and introduced an interesting (but not very reasonable) numeric measurement of self-confident and 'real' skill level of each student. So we can know which group is more likely to under-evaluate themselves. The whole python code is available in the same folder.

Data Processing

First, take a look at 'Program' and 'Gender' attributes, there are some small categories that could be grouped into other categories. For example, 'PhD Biomedical Informatics' can be grouped into 'PhD', and 'Ms in ds' into 'IDSE (master)'. In python map function could do this for us.

Then we could map F/T in our cleaned data into 1/0 and confident level into four levels: 3, 2, 1 and 0. In this way we can measure level of skill set numerically. Again, this is just for discovering some interesting facts.

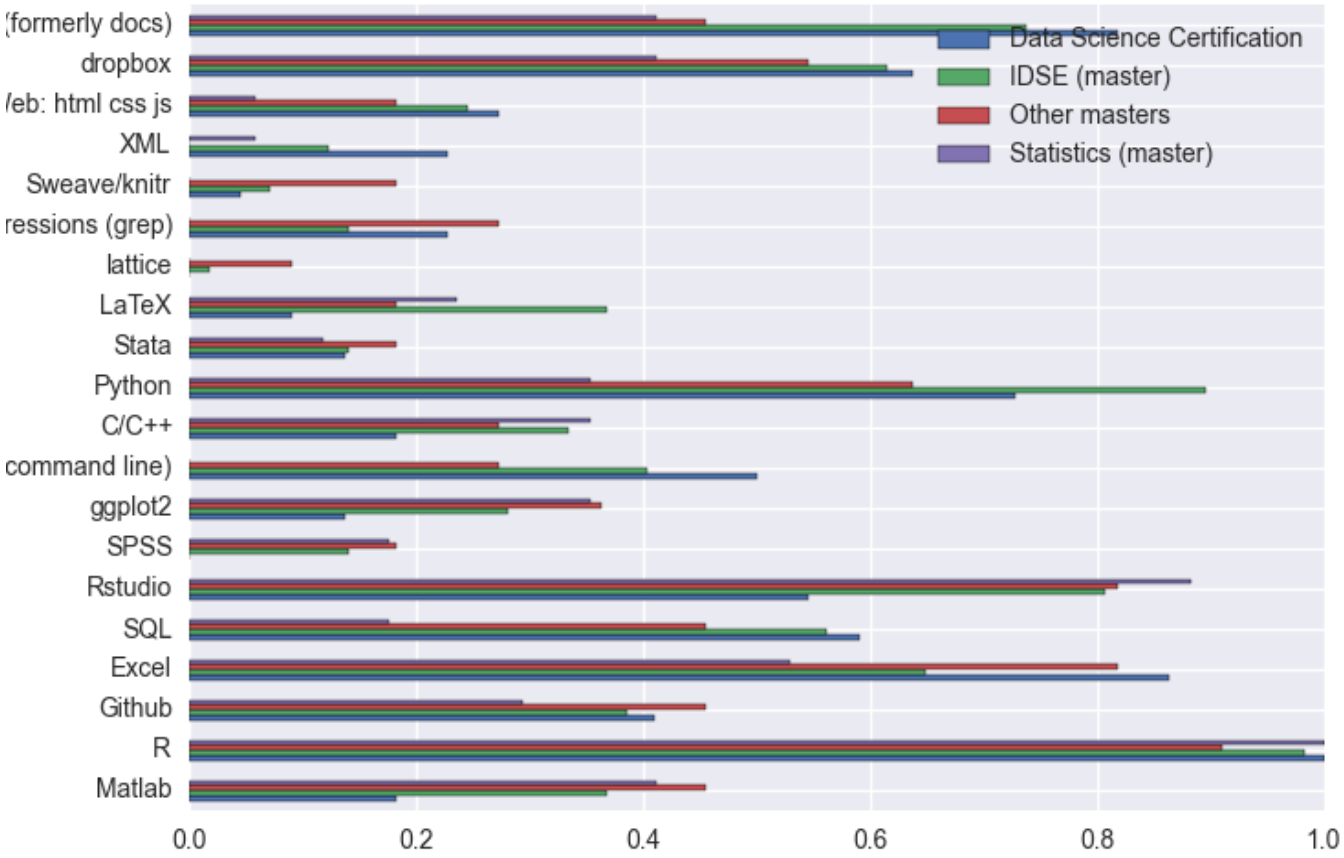
After the above transformation, we sum all skill set and confident level to get total skill score (SkillVal) and self-evaluation score (SelfVal). Adding these two columns into table. Moreover, it is intuition to set data['over_evaluation'] = data['SelfVal'] - data['SkillVal']. Therefore, we can say students who get positive numbers in 'over_evaluation' tend to be over-confident. Otherwise they are under-evaluating themselves.

Data Visualization

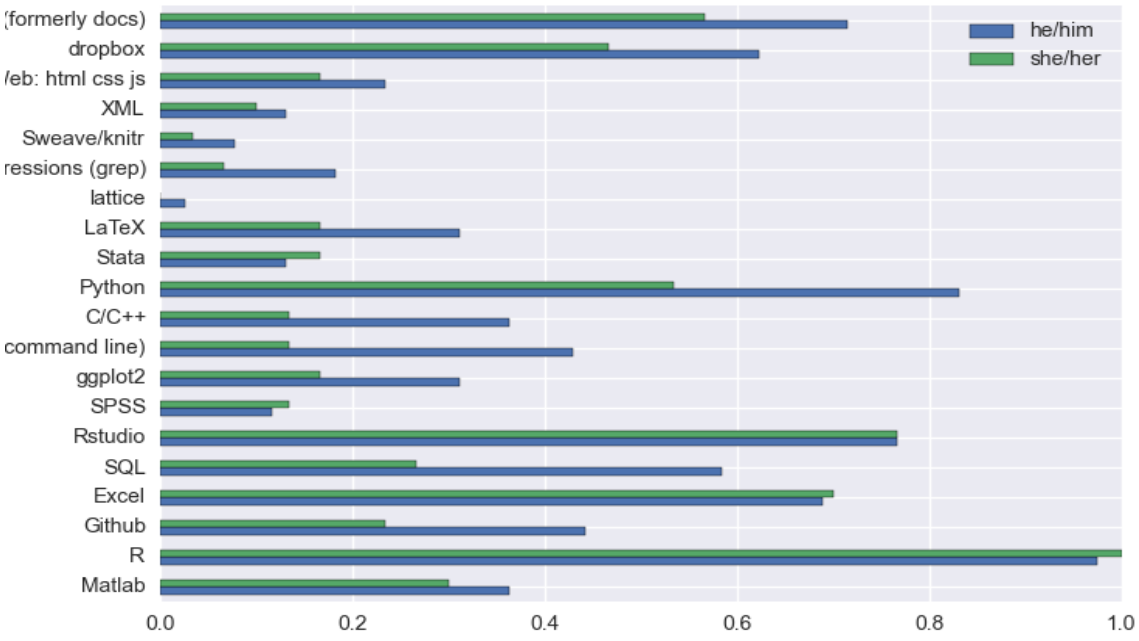
Now let us try to plot some graphics. As the result of the above data processing, we get cleaned data frame and many grouped data. Thus, it is easier to us to plot a graph with much more information.

The following plots are some plots of average knowledge level of each skill by program, gender, and waiting list.

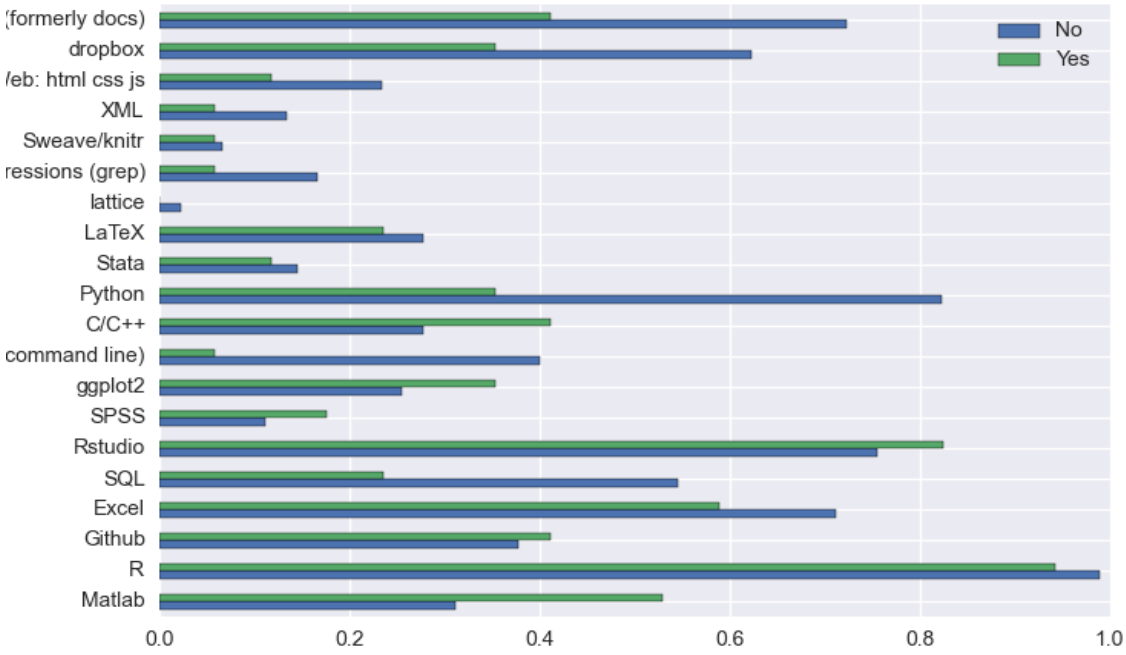
Notes that to let visualization makes more sense, we excluded some small categories whose count is less than 5.



Average score of each skill by program. We can see many info from this picture. For example: the majority uses R and RStudio before; Data Science and stats masters has stronger background in C/C++ because some of them are from China where C/C++ is a required course for them in their undergraduate. Students from Data Science Certification are more likely to have background in some industry skills such as XML, dropbox, and Excel. Stats students do not so much programming background in python, command line, css (cs skills).

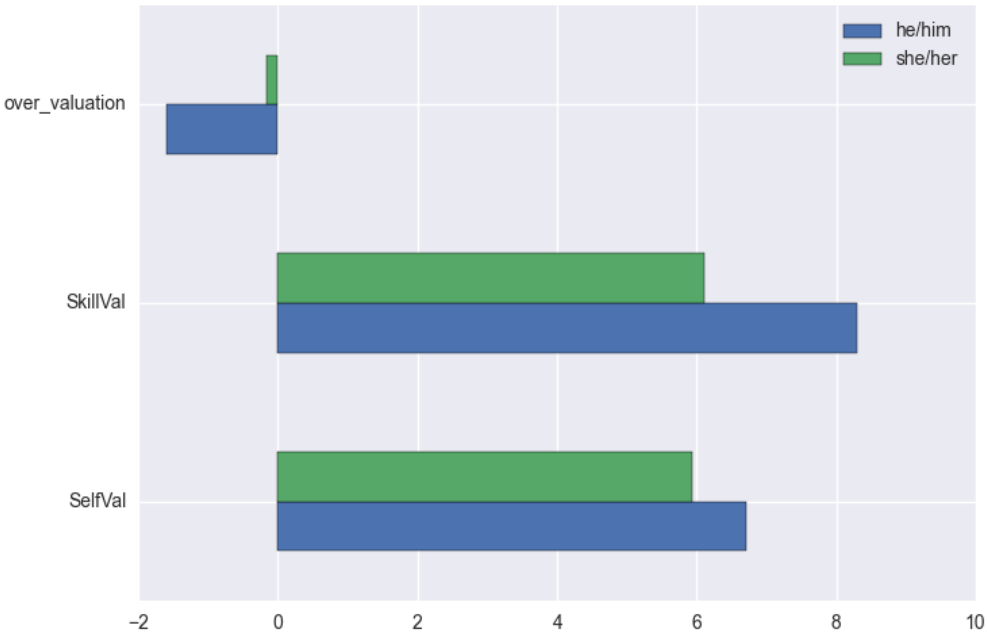


Average score of each skill by gender. Male has stronger background than female in most skills.

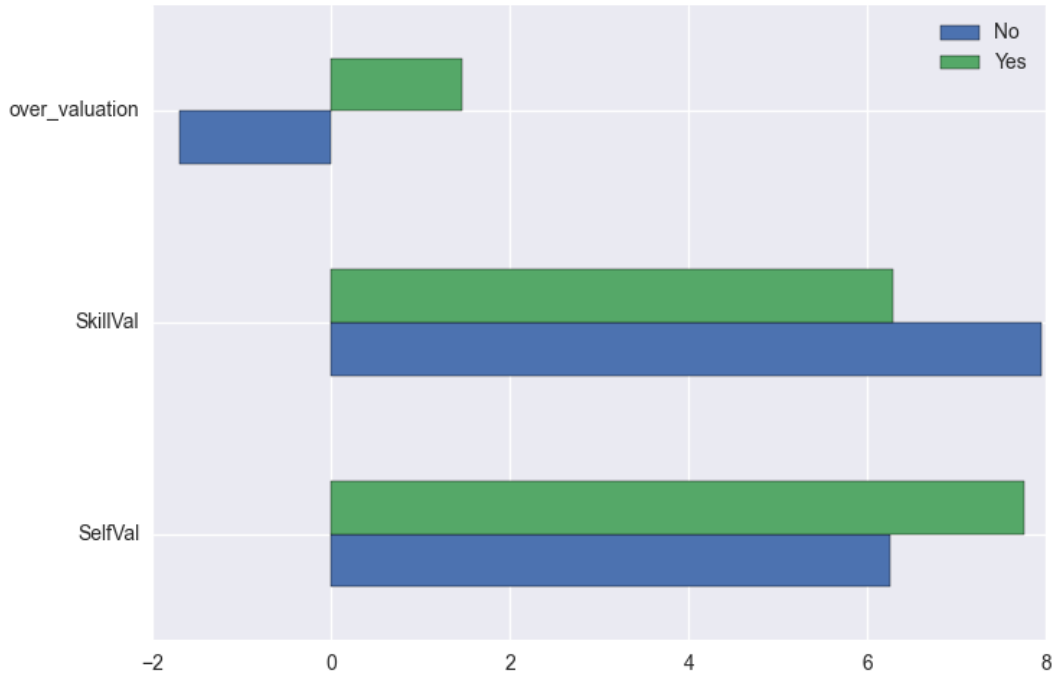


Average score of each skill by waiting list. People in waiting list are mostly from stats. So that is why they have stronger stats background but weaker programming background.

Interesting plots for self-evaluation and confident level by program. We can see that stats people are likely to over evaluate themselves but data science students are modest. Just for fun.



Plots for self evaluation and confident level by gender. It shows male intends to under evaluate themselves.



Plots for self evaluation and confident level by gender. People in waiting list are mostly from stats and here waiting list people (yes in plot) intend to be over-confident. The same result as above.