# First assignment: The Class

*February 10, 2016*

## Introduction

The first exploratory project in the Exploratory Data Analysis and Visualization focuses on the survey response dataset, acquired through the online questionnaire about statistical and technical literacy distributed among the students of EDAV class.

We want to analyze the characteristics of EDAV students in terms of 3 overall dimensions:

- Knowledge of tools
- Confidence in skills
- Preference of text editor

The data we will be working with for this assignment consists of survey responses from **114** graduate students, which are enrolled in one of the following programs:

- IDSE
- Data Science Certificate
- Statistics
- Others (include QMSS, Applied Math, Biomedical Informatics etc.)

### Cleaning the data

In order to work with clean and consistent data, we created a function called 'tinydata' that takes the *xlsx* file containing the raw responses and returns a consistent data frame used across all the analyses performed.

Tidy data transformations are stored as a script file named `tidydata.R` in our working directory that is sourced and run before we start the analysis.

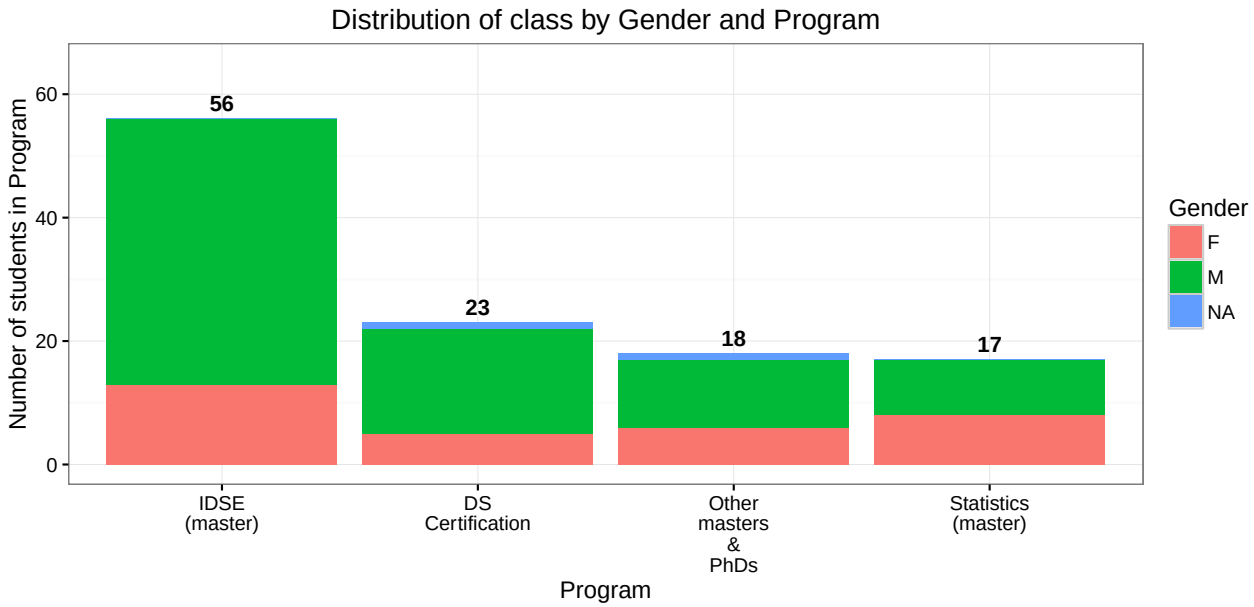
In a nutshell, tidydata does the following:

1. Removes the columns that have no observations in them.
2. Regroups the columns in a logical manner.
3. Organizes the factor levels in the following columns
    a. Waiting List Info
    b. Gender
    c. Program
    d. Text Editors
4. Splits the "Experience with Tools"" column to multiple columns with a 1/0 value depending on if the skill was reported or not. Splitting the variable makes it tidy.

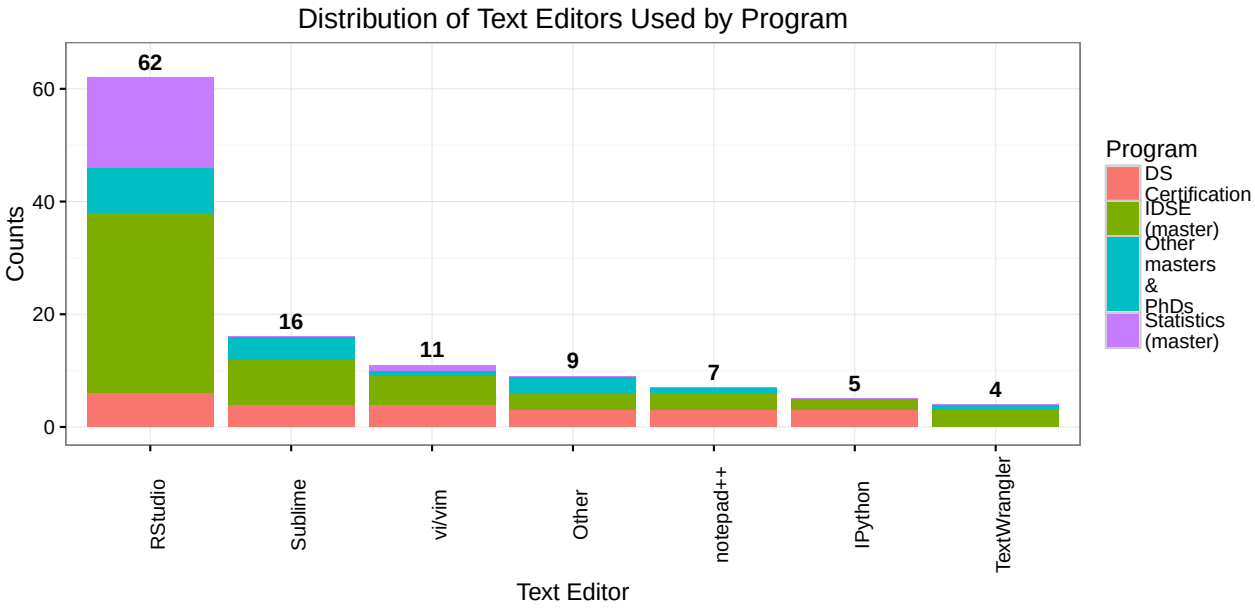
The codebook for tidy dataset is provided in the project main directory (`codebook.txt`), one can refer to it regarding the information each variable, its possible values, levels etc.

The distribution of respondents by program and gender is as follows:

```
#the source code for each plot can be found in /src/ folder in main project directory
source("src/tidydata.R")
source("src/GenderProgramPlot.R")
filename <- "raw/Survey+Response.xlsx"
df <- tidydata(filename)
GenderProgramPlot(df)
```




As it can be concluded from the chart, the men's population is considerably larger than women's in each of the programs presented on the course with relatively the same ratio across the programs. What is more, the chart allows us to judge to what extent each of the programs is presented on this course. The largest portion of students predictably comes from DSI master program since this is the core curriculum course within this master program. This block of students is followed by DSI certificate students, and the third large identified group is Master in Statistics program. However, there is 18 students more in the course whose institute affiliation is different from the programs mentioned above, varying from QMSS programs to Biomedical Informatics ones.




Another set of interesting facts can be revealed from the distribution of preferred text and code editors across the programs presented on the course. As the majority of students come from data science and statistics programs, it is quite expected that the leading code editor named in the survey is RStudio, which is the leading IDE for R development. RStudio is followed Sublime, which is one of the most widespread and very universal code editors in the market right now, and surprisingly the third place is taken by vi/vim, which is a quite sophisticated command line text editing tool, which indicates that the course accommodates significant portion of students with vast programming background, besides the overal programming literacy on the course is sustainably high. Unpredictably few people selected IPython as their preferred text editor, though it is considered as a really convenient interactive tool for Python development, which is also extremely widespread for statistical calculations.

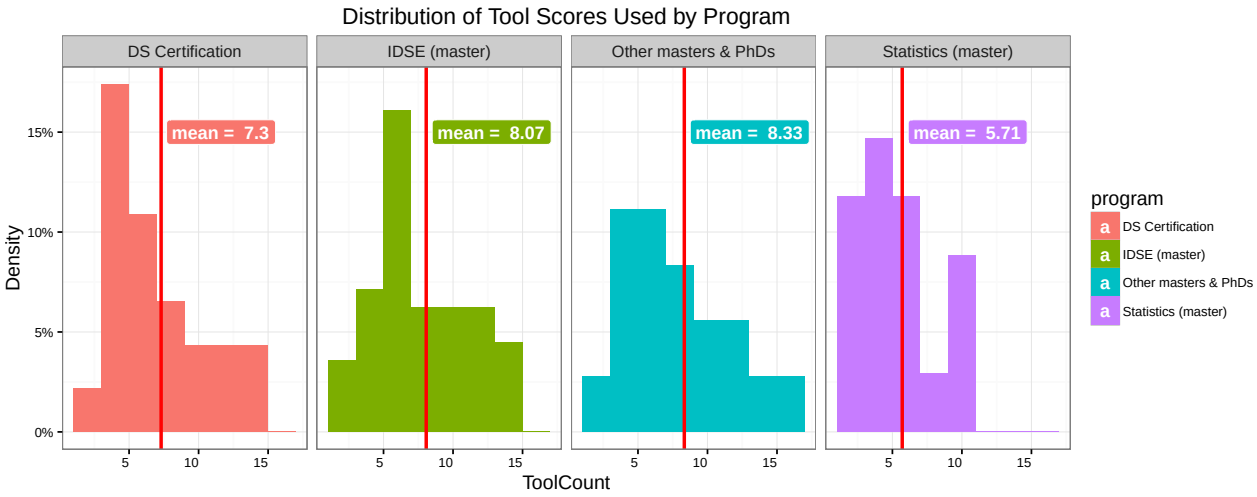
Speaking further, we would like to point out to another curious fact that can be spotted on this bar chart. Proportion of Data Science certificate students who have chosen RStudio as their preferred text editor is extremely low in both absolute and relative terms compared to the other programs and the other tools. This might suggest us the following conclusion: most certificate students come from the technical or semi-technical occupations and are exposed to writing code and developing products on the regular basis. However, this activity most probably is not limited to data analysis and statistical calculations, thus RStudio does not seem to be a natural choice for them and they stick to more universal and generally accepted tools like Sublime, vi, notepad++ etc.

We continue our analysis by exploring different toolkits that students from different programs pick to be using.

# Tools

As we learned in the previous chapter, the respondents come from very diverse backgrounds and, as such, have learned and been exposed to very different sets of tools.

The questionnaire provided a list of **20 possible choices** to pick up from, corresponding to different tools, programming languages and frameworks, and each student was asked which of these he or she was comfortable working with. The following plot shows the distribution of the number of data-related tools students from each derived program category claimed to use confidently.



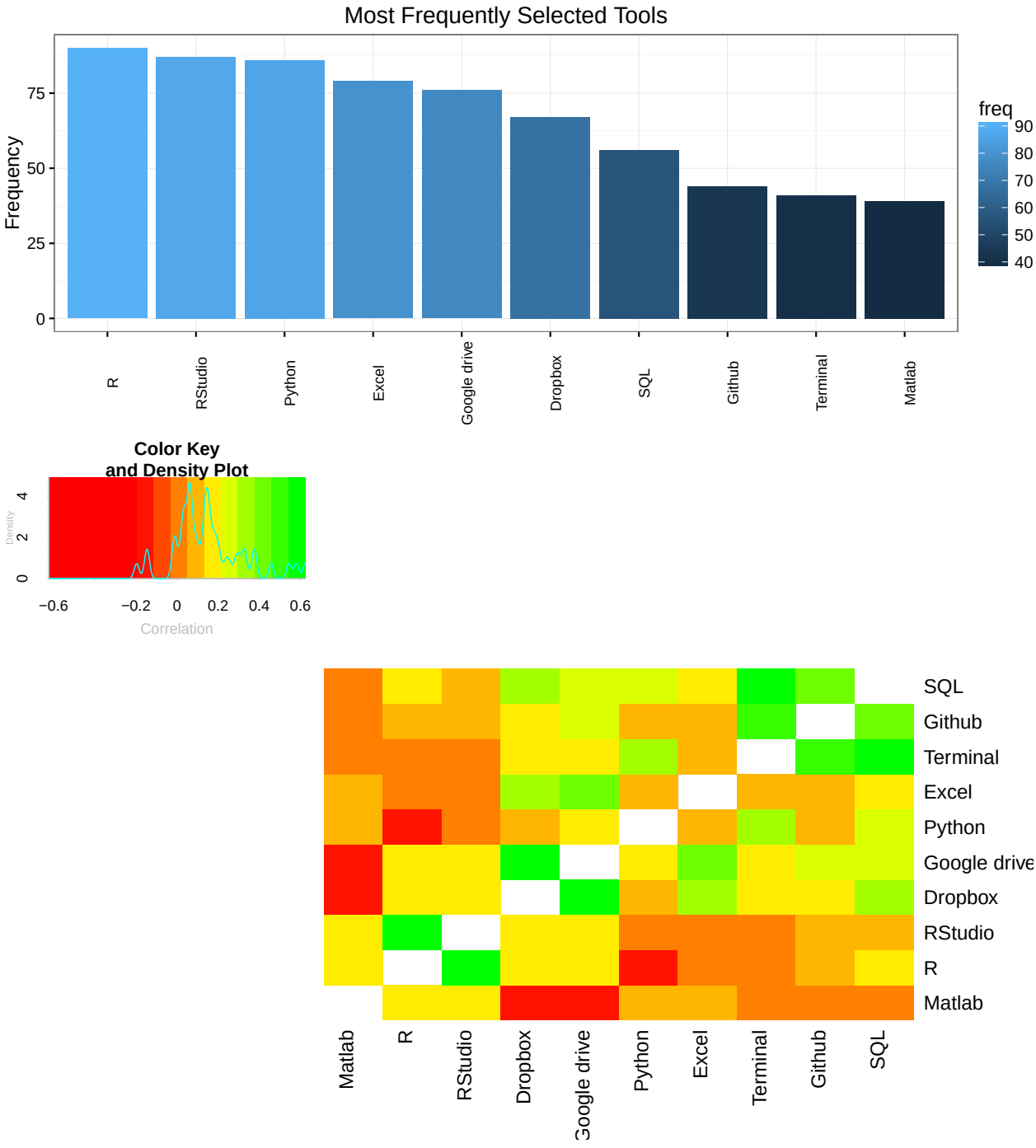

Taking into account the number of people in each group we can say that with certain level of confidence we can treat each of these four distributions as normal. And taking a look at the means of each distribution plotted above the bars, we can extract some potentially valuable knowledge:

- Master students on Data Science program on average have a knowledge of bigger amount of skills considered to be data-related compared to cirtificate students. And as we remember from the text editor preference exploration, it corresponds perfectly with the fact, that certificate students have a solid level of programming literacy, though they most probably come from different fields of IT industry and their work stack includes different programming languages and frameworks.
- Group of students from all other programs tend to have the highest average score, which is really interesting, though might be an indicator of bigger academic experience among these students, hence broader range of used tools in data analysis. As we remember PhD students were aggregated to this category and they can be those outliers, who drive the sample average to bigger values.
- Students from masters program in Statistics demonstrate the lowest score. This reveal can potentially correspond to the fact that Statistics program in CU lacks practical exercises and most of their time students do theoretical *(say, "paper-based")* studies and exercises and do not have substantial experience with full stack of data science programming tools.

All these conclusions might be purely speculative, however these suggestions might fuel further hypotheses and experiments designed to prove or disprove these hypotheses.

The next question we sought to answer was whether there are correlations between different sets of tools. In order to visualize this, we created a heatmap showing the correlation matrix for the top 10 tools.

At first we filtered out top 10 frequently mentioned tools, as demonstrated on the chart below. Then this set of tools was used to build a pair-wise correlation heatmap, which is also provided below. The point was to identify the tools that compliment each other and typically are mentioned together.




As we can see, SQL, Github and Terminal have the strongest correlations, meaning that people that tend to know one of these tools, on average will know how to use the other two with a higher probability than the average student.

- This result was somewhat expected. In general people with a stronger CS background will tend to know these three programs.

Interesting insights can also be observed by looking at the negative correlations seen between Python-R and between Matlab- Dropbox and Matlab-Google Drive.

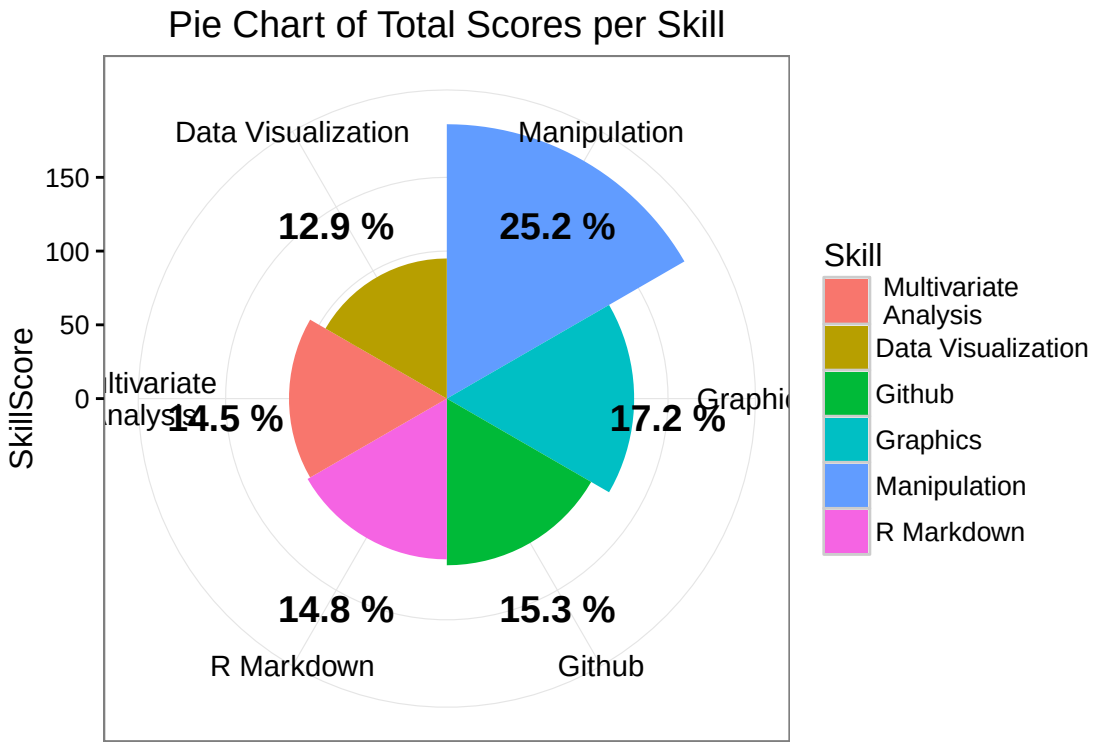
- We are not surprised to see a negative correlation between R and Python. In general, they can be regarded as "substitute" tools, so people can work with one without ever having to learn the other.
- The negative correlations between Matlab and Dropbox/Google drive are harder to explain. One potential explanation is that people that didn't know Matlab felt the need to fill in more tools, and so inputted dropbox or google drive.

## Skillsets

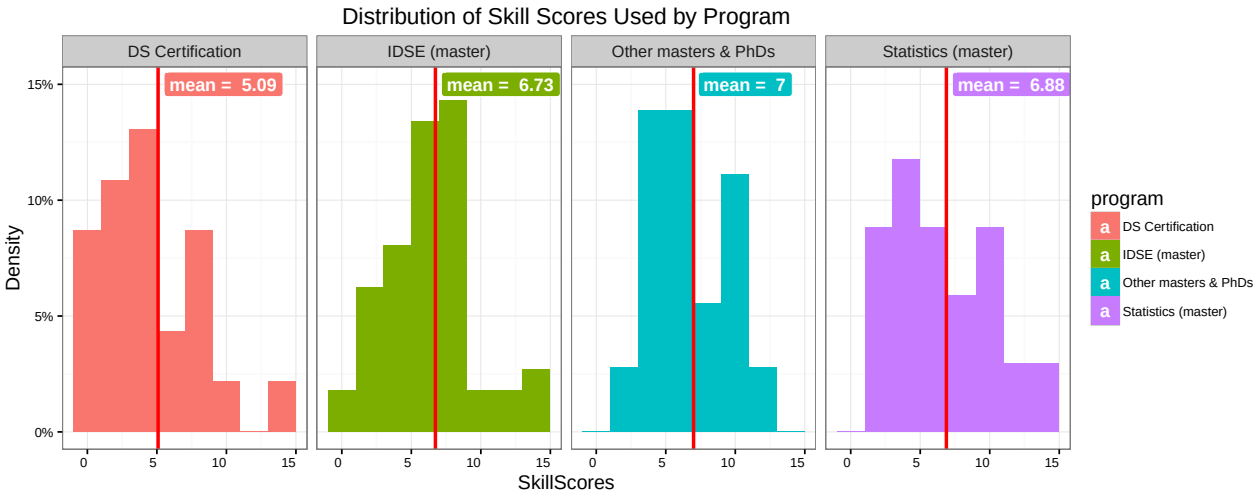
The next section of our work is dedicated to understand the skillsets that were reported by respondents.

The survey asked respondents to grade their level of confidence in 6 different dimensions which we call "Skills". The possible responses were "None", "A little", "Confident", and "Expert". As mentioned before, we assigned values from 0 to 3 to these levels to be able to aggregate data and compute averages. It is important to mention that this score is not perfect for several reasons, starting with the fact that these are self reported levels of confidence, and equal weight is being assigned to each skill.

The first exploratory analysis we did was to compute aggregated levels of confidence (called scores from here onward). The distribution of scores by skill looks as follows:



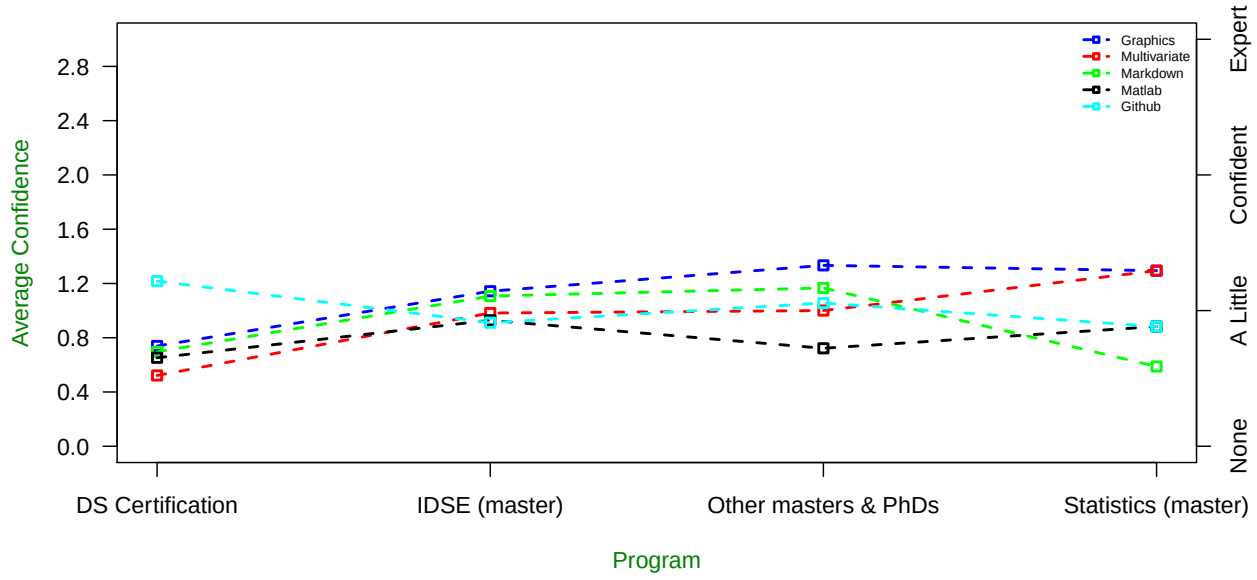

Not surprisingly, we see that multivariate analysis has the highest overall confidence level. Now, we wish to have a bit more insight about how these confidence levels look when grouping by masters.

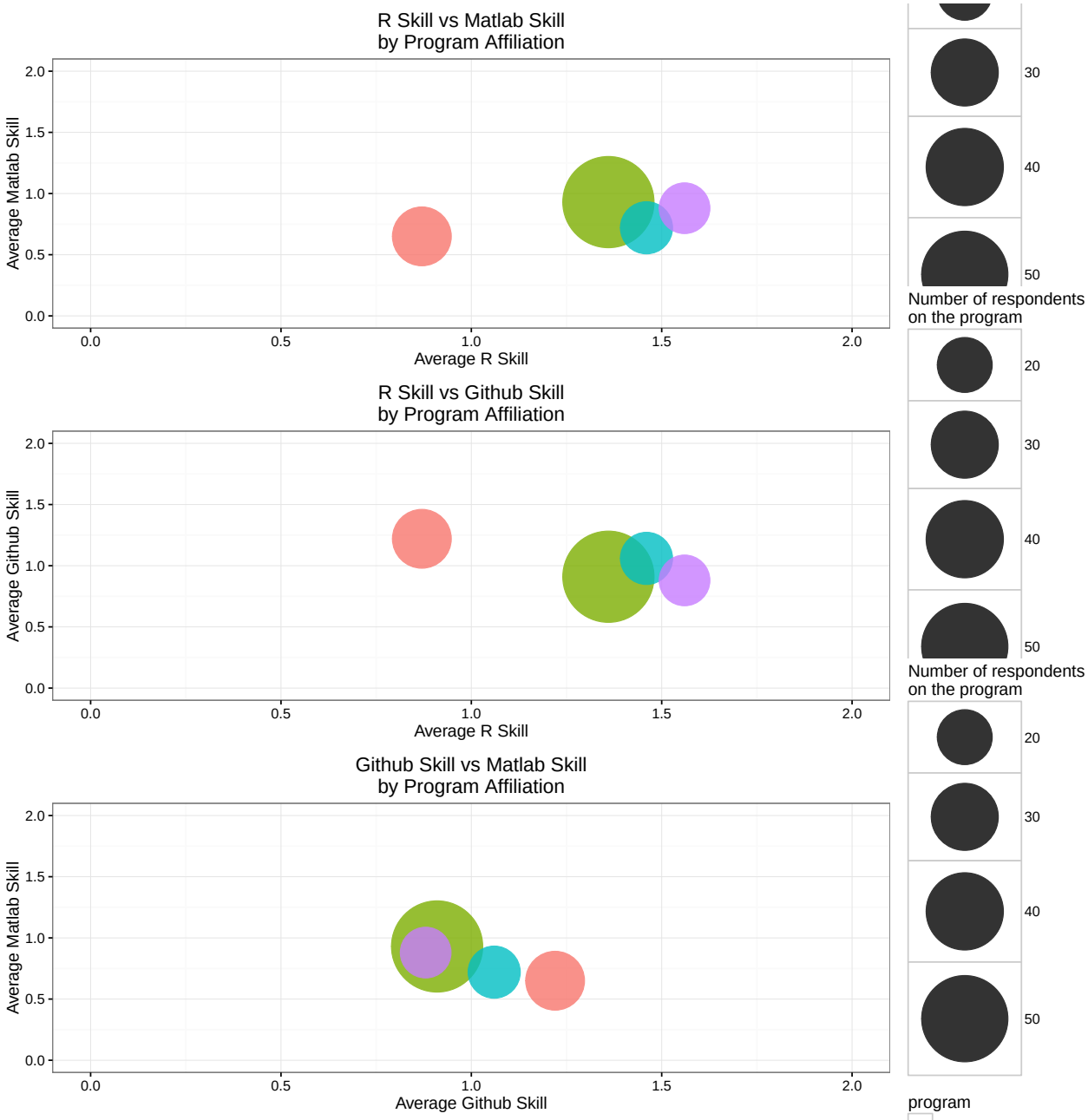



## VERIFY next statement when data about average is available

Interestingly, Data Science Certificates and Statistics masters students seem to have the lowest level of confidence in their skills.

In order to understand which masters are confident with which skills, we took the analysis one step further to show with a plot this level of detail:






## ADD A SUMMARY OF FINDINGS

## Text editor preference (NOT SURE IF WE WANT TO GO INTO THIS )