EDVA Project I: Course Students Analysis

Manuel Rueda

February 11, 2016

Introduction

On this project we look into the EDAV course intro survey data to explore the characteristics of enrolled students and their proficiency with a number of tools useful for the course. First we will start by performing exploratory analysis on the data, analyzing the distribution of each of the variables. Asides from the skill-related information, the survey provides us with some demographic data, such as the gender of each of the students, which program they come from and their enrollment status for the course (enrolled vs. on waitlist). Using the demographic data we are able to generate conditional analysis on the skills and tools data, to identify if there are significant differences between the populations. Finally we dive deeper into the data by generating a skill ranking system for each of the students and correlation, cluster and PCA analysis on the distributions of tools and skills.

Data Cleaning

Before starting to analyze the data, first we are required to clean it and standardize for ease of access. Below we outline the steps taken for this, with specific available at the code file.

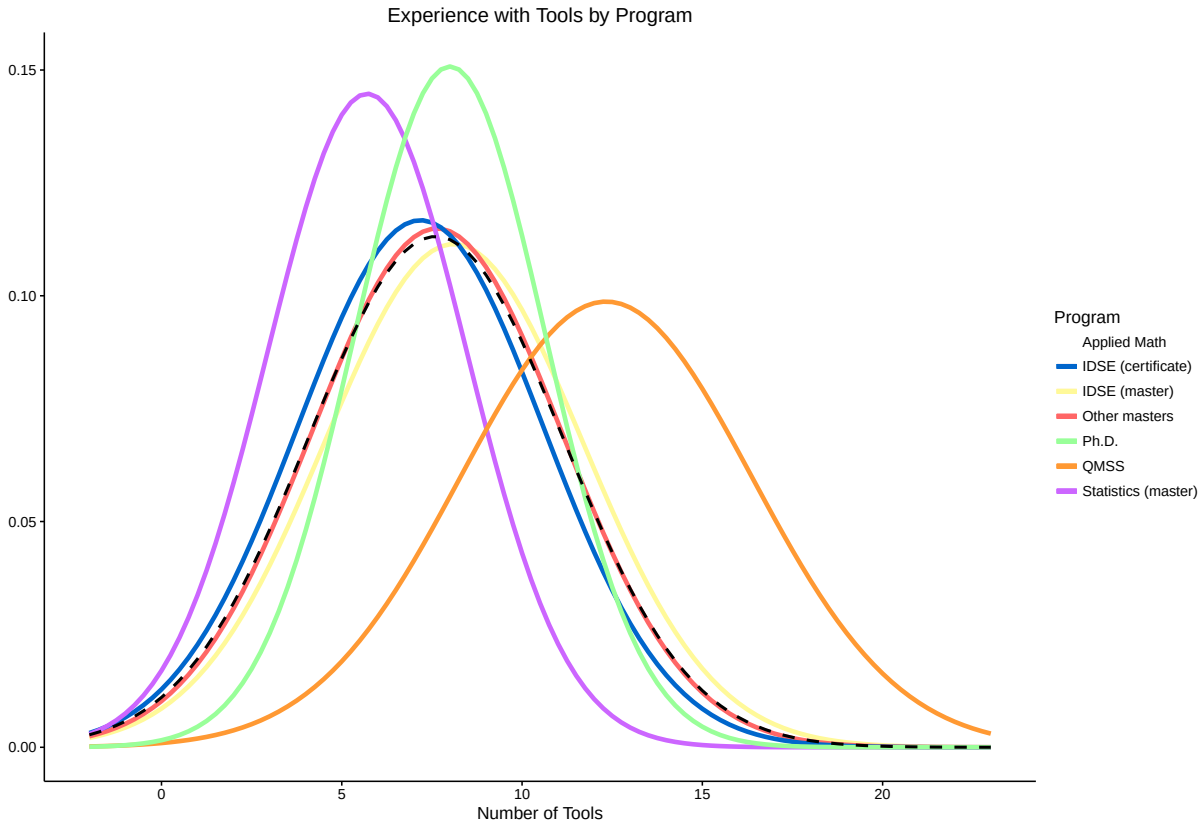
1. The input file contains a number of rows that provide no information (i.e. all *NA* values). We will proceed to delete these.
2. The second column, *Program*, contains distinct entries that refer to the same program, such as "IDSE (master)" and "master in ds". We will standardize this.
3. The third column, *Experience with tools*, is in the form of a list within each row. For purposes of our analysis we need to parse the column into a number of binary variables that will indicate wether or not a given student has experience with a particular tool.
4. Similar to point *2*, the *Preferred Editor* column contains redundant entries. We proceed to standardize and clean these up.
5. For *Gender*, two entries did not correspond to male or female. We have randomly assigned one of the 2 genders to them (given the large number of observations the impact of this transformation is negligible).

Below is a snapshot of how the final working data frame looks like.

```
##   id Waitlist            Program R - Data Manipulation  Gender
## 1  1       No      IDSE (master)             Confident she/her
## 2  2       No      Other masters             Confident  he/him
## 3  3       No IDSE (certificate)                Expert  he/him
##   Pref. Editor R - Graphic Skills R - Adv. Multivariate Analysis Skills
## 1      RStudio          Confident                                A little
## 2      RStudio           A little                                A little
## 3         Atom          Confident                               Confident
##   R - Reproducibility Skills Matlab Skills Github Skills R Excel SQL
## 1                  A little      A little          None 1     1   1
## 2                  A little      A little      A little 1     1   0
## 3                    Expert     Confident     Confident 1     1   1
```

```
##   RStudio ggplot2 Python Stata Dropbox Google Drive RegEx Github Shell
## 1       1       1      1     1       1            1     0      0     0
## 2       1       1      0     0       1            1     1      0     0
## 3       1       1      1     0       1            1     1      1     1
##   LaTeX Sweave/knitr XML Web C/C++ Matlab SPSS lattice
## 1     0            0   0   0     0      0    0       0
## 2     0            0   0   0     0      0    0       0
## 3     1            1   1   1     0      0    0       0
```

Exploratory Data Analysis



The graph above shows the normal curves of the number of tools that students have experience with, based on what program they are in. The black, dashed line is the normal curve for the number of tools that students have experience with for the entire class, not taking into account what program they are in. You'll notice that the normal curve for the Applied Math students is missing. This is because there is only one Applied Math student according to the survey data. This student has experience in 5 tools, which, compared to the averages of the other programs, is the lowest.

At first glance, you might notice that the orange normal curve, which is for the QMSS students, is significantly farther to the right than the other curves, indicating that QMSS students, on average, have experience with the most amount of tools (12.33333). The QMSS curve is also the widest, suggesting that the range of number of tools that students have experience with is large.

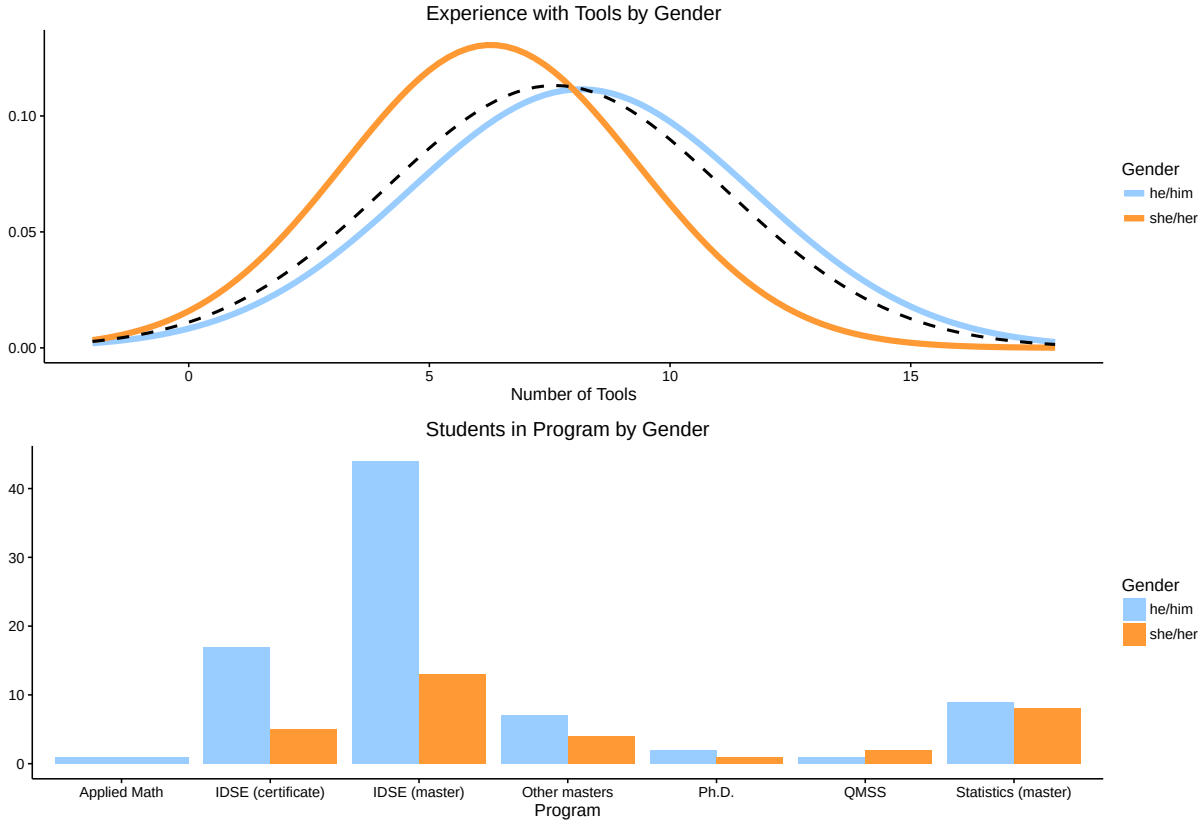
The purple normal curve, for the Statistics (master) students, is the farthest to the left, indicating that those students have experience with the least amount of tools on average (excluding the Applied Math students). This curve is also tall and narrow, suggesting that the number of tools that these students have experience with does not stray too far from the mean. It makes sense that the statistics and math students have

experience with less of these tools on average compared to programs like IDSE (master), IDSE (certificate), and QMSS, considering that the latter programs are likely more focused on computer programming.

The yellow, red, and blue normal curves, for the IDSE (masters), Other Masters, and IDSE (certificate) students, respectively, look to be clustered in the same area, and in fact, their means vary by less than 1 and standard deviations differ by less than .2. Considering the fact that almost 80% of the class is in one of these 3 programs, it should come as no surprise that the black, dashed normal curve representing the class as a whole is also in this cluster. As one might expect, the IDSE (master) students have experience with more tools than the IDSE (certificate) students on average.

The mint normal curve for the Ph.D. students is the tallest and most narrow, indicating that compared to all of other programs, the number of tools that these students have experience with varies the least. The Ph.D. average is larger than the class average, which I assume can be attributed to the fact that they are working towards getting the highest-level degree of anyone in the class.

Experience With Tools by Gender



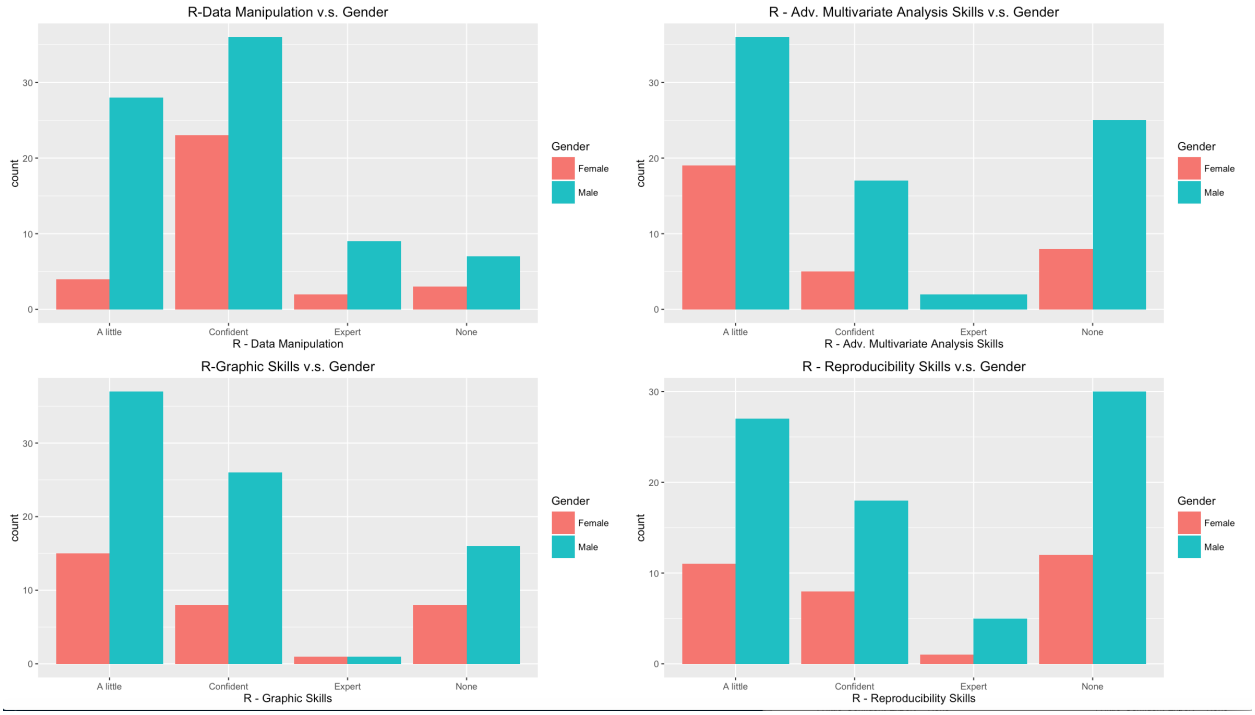
The graph above plots the normal curves of the number of tools that students have experience with, based on their gender. The black, dashed line is the normal curve for the number of tools that students have experience with for the entire class, not taking into account gender.

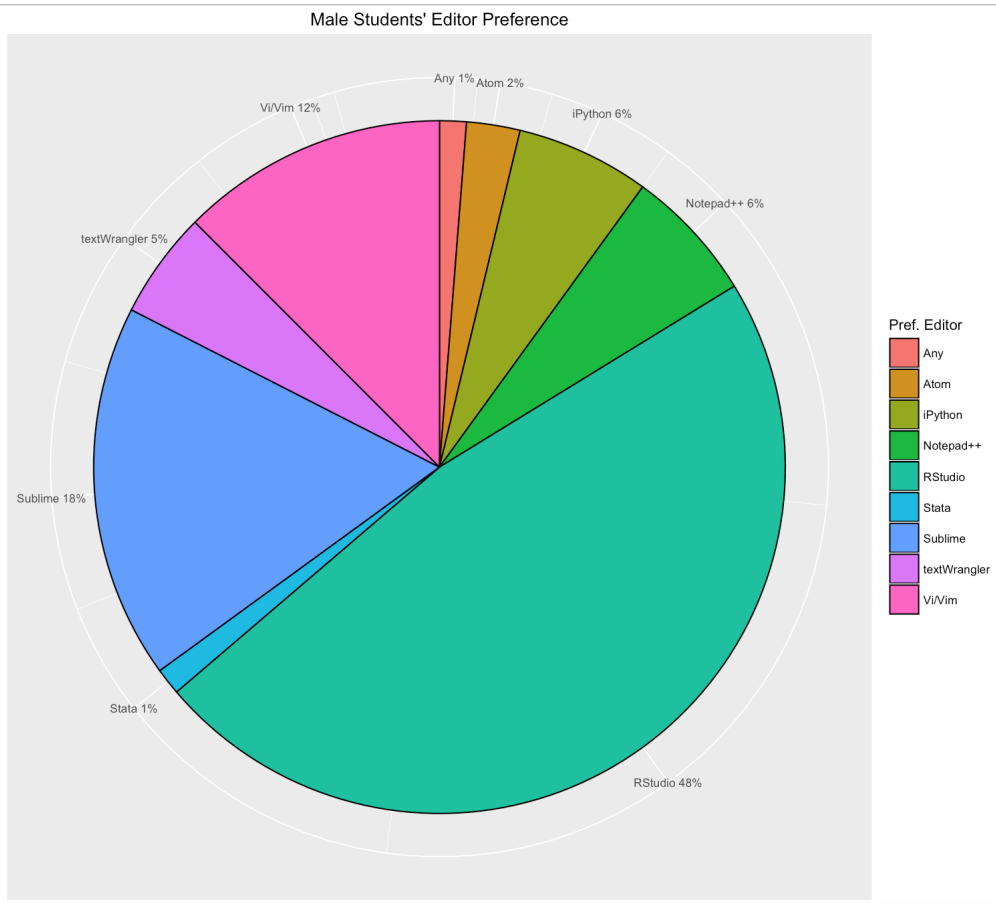
The normal curve for the entire class is similar to that of the male students, and certainly more similar to the male students than the females. This is likely because there are more than double the amount of males in the class than there are females. Also, the curve for the females is taller and slightly skinnier, which means the amount of tools that females have experience with varies less than that of the males.

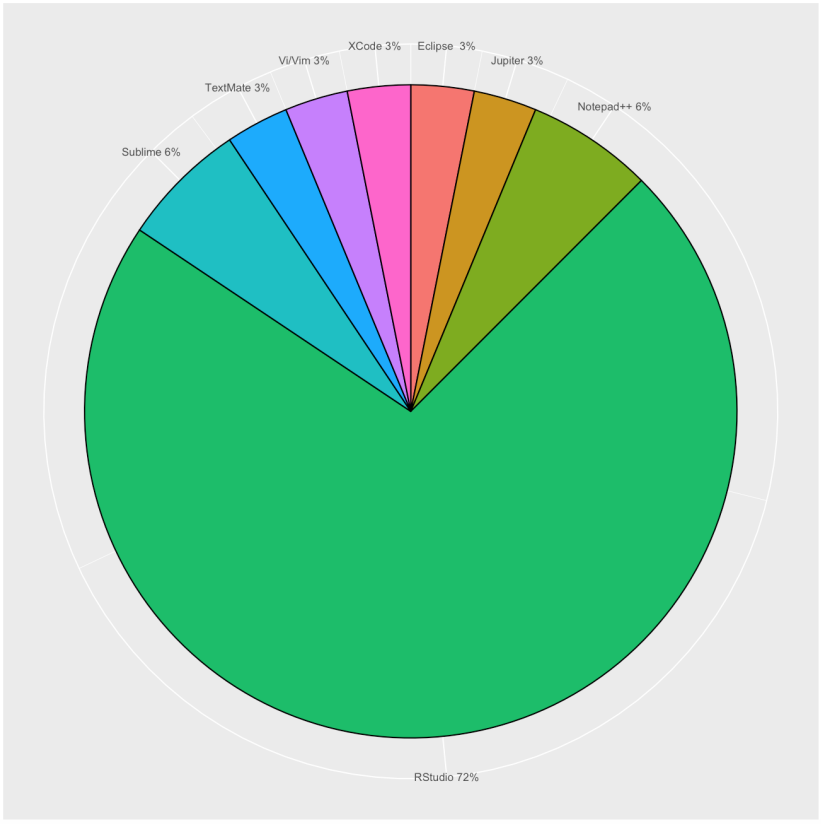
The difference in averages between the genders is most likely do to the programs in which the students are enrolled. There is more than 3 times as many males in the IDSE (master) program than there are females,

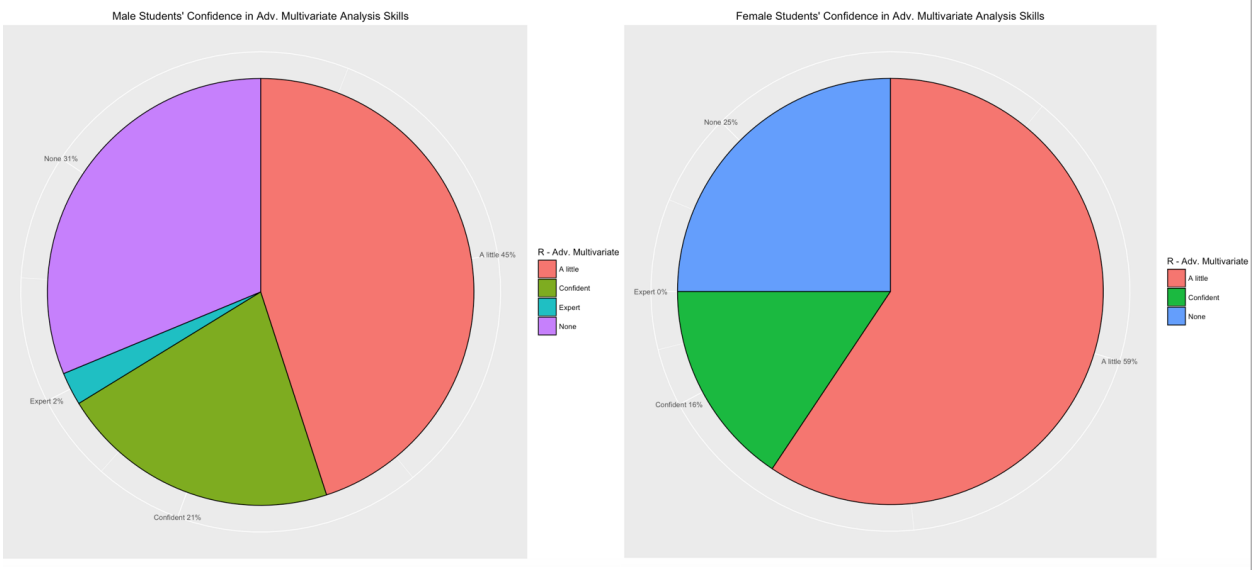
and that program had the second largest average for the number of tools that students had experience with. Further, the program that had the second most amount of females in it was the Statistics (master) program, which had the second lowest average for the number of tools that students had experience with.

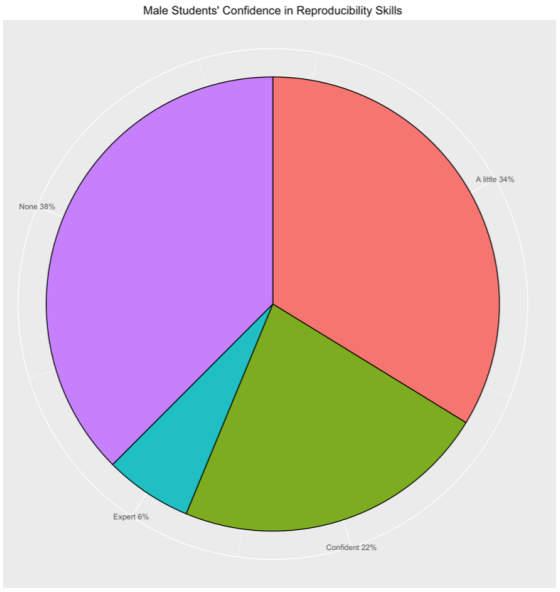
Conditional Distributions

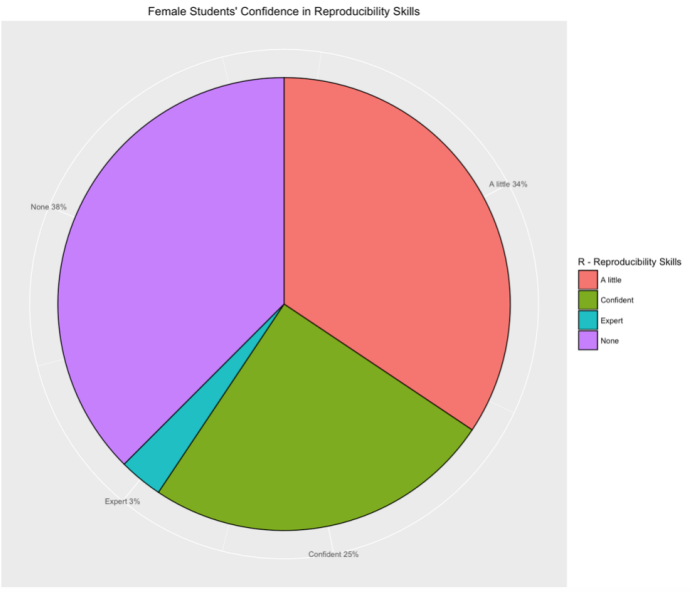


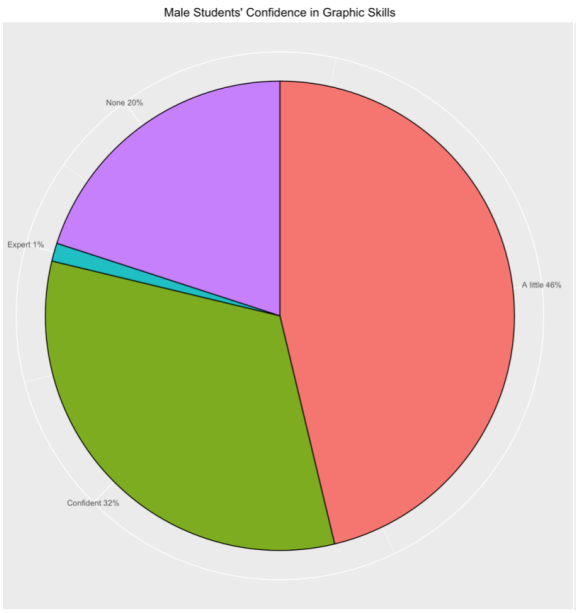

Male Students' Editor Preference
Any 1%
Atom 2%
iPython 6%
Notepad++ 6%
RStudio 48%
Stata 1%
Sublime 18%
textWrangler 5%
Vi/Vim 12%
Pref. Editor
Any
Atom
iPython
Notepad++
RStudio
Stata
Sublime
textWrangler
Vi/Vim

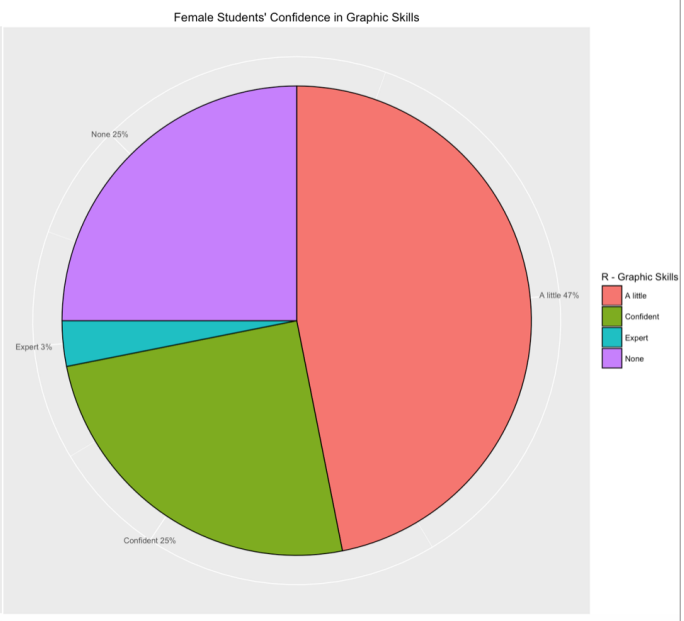

Female Students' Editor Preference
XCode 3%
Eclipse 3%
Jupiter 3%
Vi/Vim 3%
TextMate 3%
Notepad++ 6%
Sublime 6%
RStudio 72%
Pref. Editor
Eclipse
Jupiter
Notepad++
RStudio
Sublime
TextMate
Vi/Vim
XCode

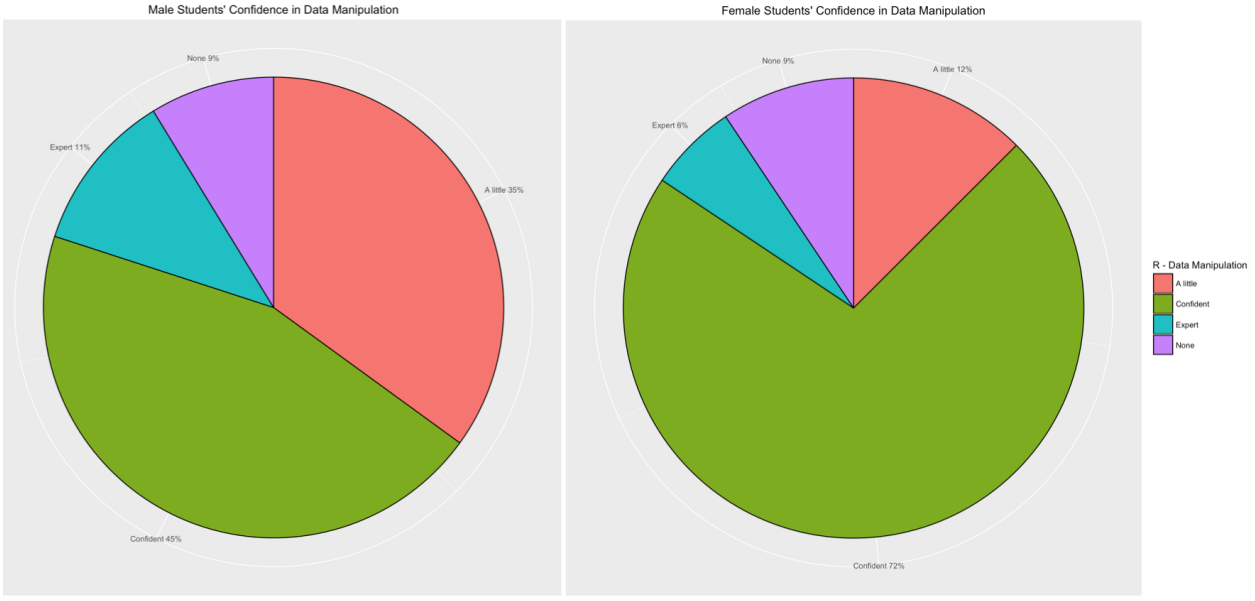

Male Students' Confidence in Adv. Multivariate Analysis Skills
None 31%
A little 45%
Expert 2%
Confident 21%
R - Adv. Multivariate
A little
Confident
Expert
None
Female Students' Confidence in Adv. Multivariate Analysis Skills
None 25%
Expert 0%
A little 59%
Confident 16%
R - Adv. Multivariate
A little
Confident
None


Male Students' Confidence in Reproducibility Skills
A little 34%
Confident 22%
Expert 6%
None 38%


Female Students' Confidence in Reproducibility Skills
A little 34%
Confident 25%
Expert 3%
None 38%
R - Reproducibility Skills
A little
Confident
Expert
None


Male Students' Confidence in Graphic Skills
A little 46%
Confident 32%
Expert 1%
None 20%


Female Students' Confidence in Graphic Skills
A little 47%
Confident 25%
Expert 3%
None 25%
R - Graphic Skills
A little
Confident
Expert
None



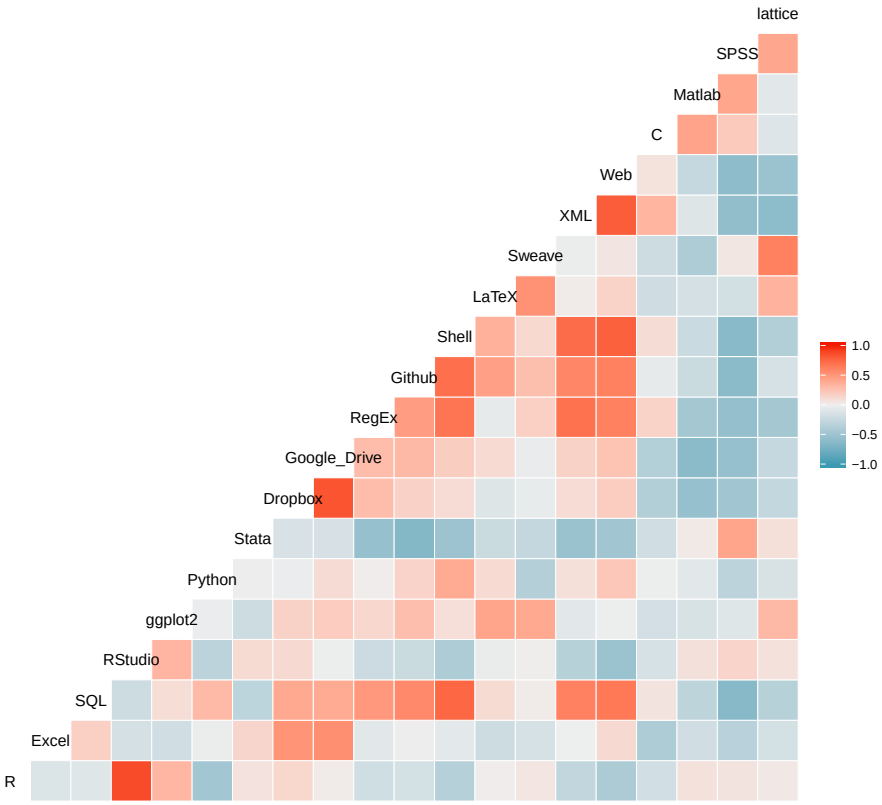
Correlation, Cluster Analysis and PCA on Skills

For this part of the analysis, we will focus our attention on the *Tools* that each of the students listed as being comfortable with. There are a total of 20 different reported tools, an we would like to investigate if some of these tend to appear in groups. An obvious example is to expect that *R* and *RStudio* are generally reported together, while *Web* and *Matlab* probably are not so closely related.

The list of reported skills are:

```
##  [1] "R"            "Excel"        "SQL"          "RStudio"
##  [5] "ggplot2"      "Python"       "Stata"        "Dropbox"
##  [9] "Google Drive" "RegEx"        "Github"       "Shell"
## [13] "LaTeX"        "Sweave/knitr" "XML"          "Web"
## [17] "C/C++"        "Matlab"       "SPSS"         "lattice"
```

As a first step, we begin by analyzing the correlation across tools. The observed relationship is presented on the matrix below.



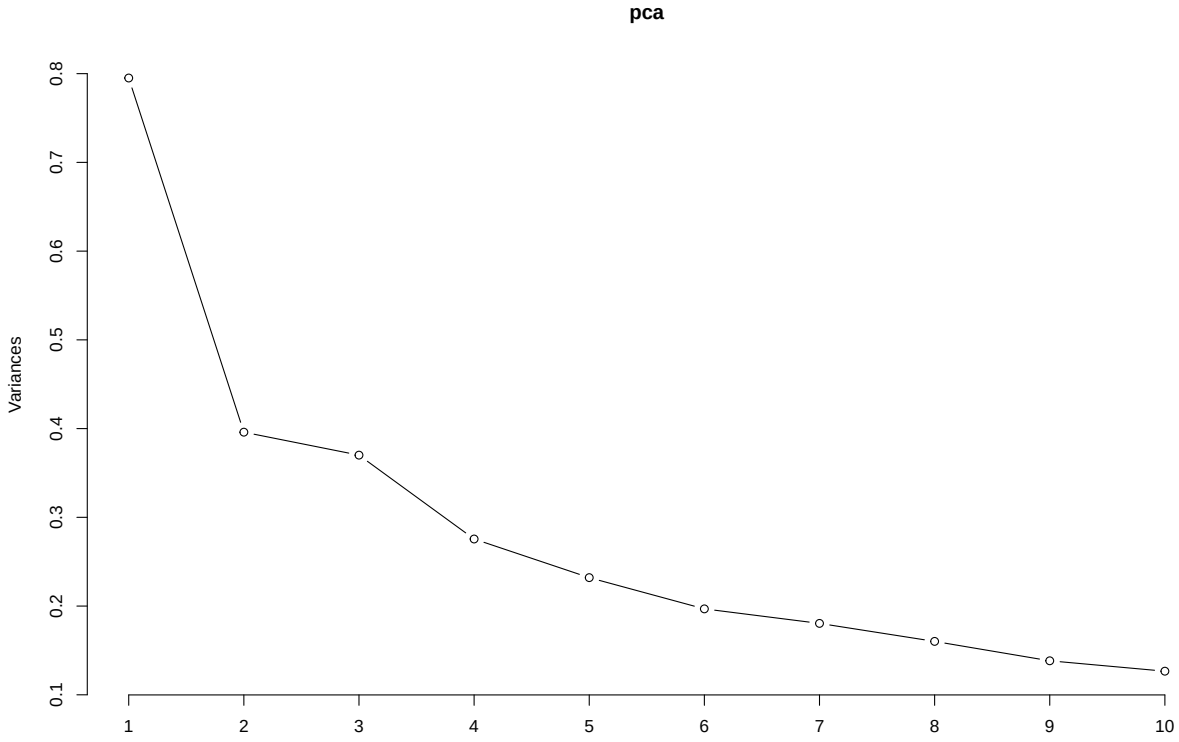
Notice that skills in R and RStudio are very strongly correlated - the two applications are very similar and they indicate the same skill in R; Dropbox and Google Drive are also highly correlated because they are both storage applications; Web and XML have strong correlation because they are highly related skill sets; Shell and Github are correlated because people who use Github tend to write command lines in Shell; it is interesting to see that SQL and Shell also have strong correlations too. In addition, there are skills that are negatively correlated. For instance, Web, XML, Shell, Github, and SQL all have strong negative relationships with SPSS. It makes sense because SPSS is a statistical package for social science - people who know SPSS tend to come from a social science background and hence don't know much about web development and data science. Further, Stata and Github have strong negative correlation because data scientists and statisticians who use Github for storage and collaboration are likely to be expert in R and hence would not use STATA. In sum, the correlation matrix gives us a clear view of which skills are highly related. It also gives us an idea of which skills are transferable to one another.

The correlation matrix gives us a good idea of how the skills relate pairwise, but it would be interesting to see how each of them relate to the rest, giving us a broader picture. Building a hierarchical binary dendrogram is useful for this.

This picture provides us some interesting insights:

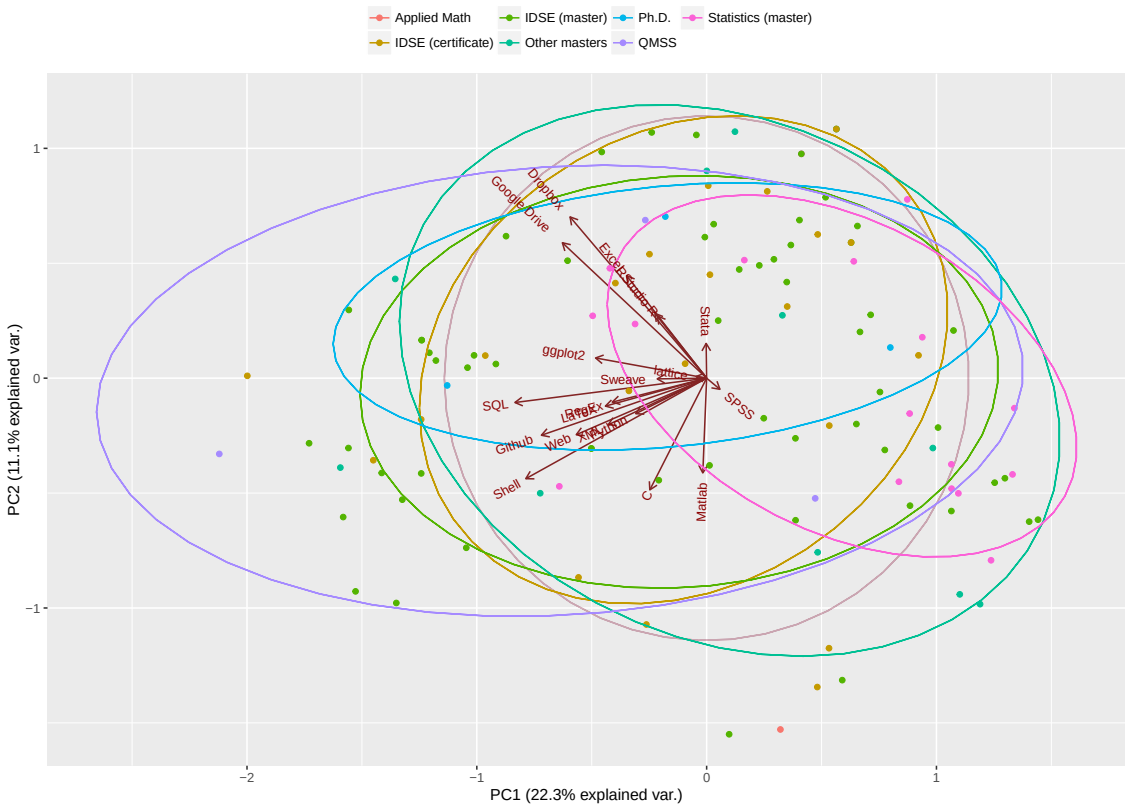
1. As expected, *R* & *RStudio* are closely related, followed by *Python*. These tools can be considered the core of the Data Scientist.
2. *Excel*, *Dropbox* and *Google Drive* also cluster together. These can be considered the least technical of the skills.
3. *XML* and *Web* go together, which makes sense considering XML and HTML are basically the same language.
4. *Stata*, *SPSS*, *Sweave* & *Latice* are tools closely related to each other. Generally these are more associated to the social sciences, or the QMSS program.

As a next step, we will perform a PCA analysis on the tools data. The idea is to reduce the dimensionality of the data by finding which tools are similar enough to be grouped with each other, so a visual representation of the data becomes possible (visualizing 20-dimensional data is a complex task). First we derive 10 PCA components, and investigate how much variance each of them explains.




```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     0.8917 0.6293 0.6083 0.52496 0.48171 0.44366
## Proportion of Variance 0.2235 0.1113 0.1040 0.07746 0.06522 0.05533
## Cumulative Proportion  0.2235 0.3348 0.4388 0.51628 0.58151 0.63683
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.42482 0.40023 0.37192 0.35582 0.34504 0.32006
## Proportion of Variance 0.05073 0.04503 0.03888 0.03559 0.03346 0.02879
## Cumulative Proportion  0.68756 0.73259 0.77147 0.80706 0.84052 0.86932
##                           PC13    PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.3167 0.29622 0.27041 0.24260 0.23268 0.21046
## Proportion of Variance 0.0282 0.02466 0.02055 0.01654 0.01522 0.01245
## Cumulative Proportion  0.8975 0.92218 0.94273 0.95927 0.97449 0.98694
##                           PC19    PC20
## Standard deviation     0.18526 0.11015
## Proportion of Variance 0.00965 0.00341
## Cumulative Proportion  0.99659 1.00000
```

The table shows us that 30% of the variance in the data can be explained by 2 components, and almost 70% by 7 of them. For purposes of visualization, we will take only 2 components and see how the skills can be represented in space, while we also investigate their different distributions across academic programs.



Again we see some interesting patterns: the least technical skills (*Excel*, *Dropbox*, *Google Drive*, etc.) are grouped on the upper left quadrant, while the more technical ones on the lower left. *SPSS* in particular has an X component with a distinct direction than the rest of the tools, indicating a dissociation with them.

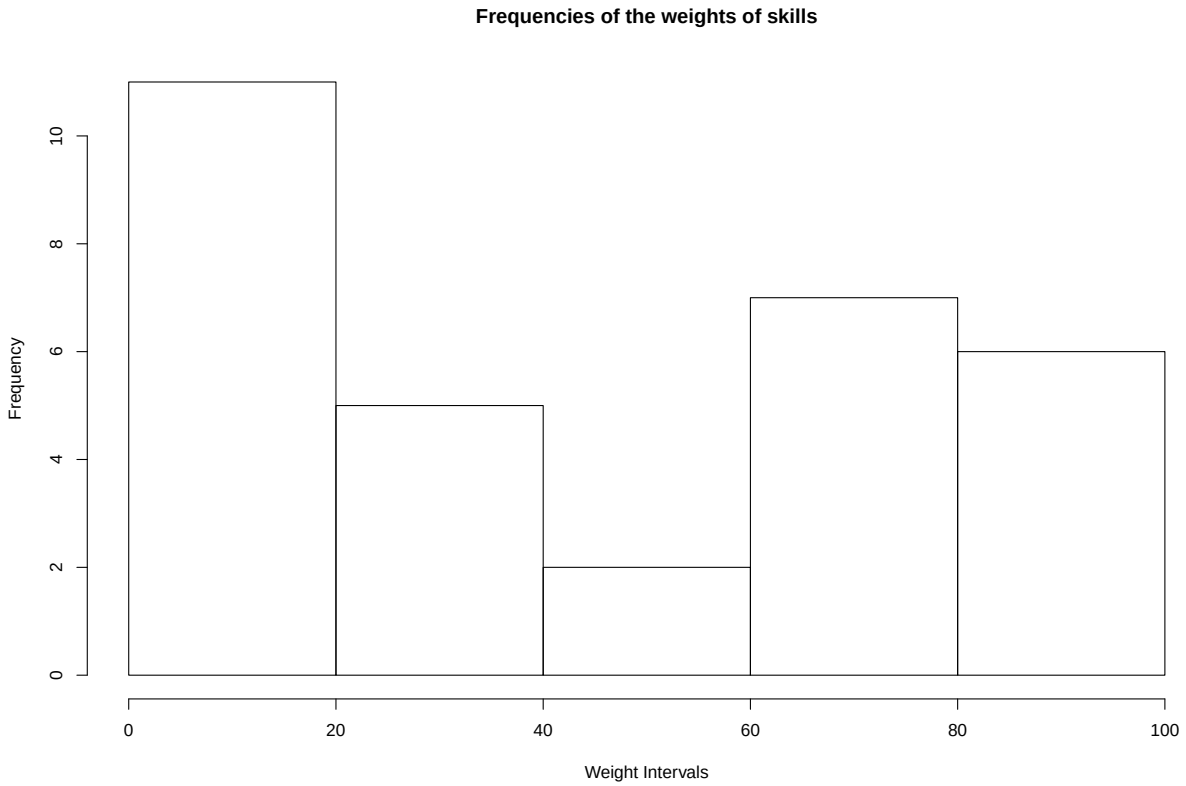
Related to the programs, we see that all of them are very similar, with 2 interesting exceptions: 1. The *Statistics (master)* group is shifted towards the right, indicating a different familiarity with tools. This group can be associated with the *SPSS* vector. 2. The *QMSS* group is the broadest one, implying that their skills vary significantly form student to student. This is something we already observed on previous analysis from our study.

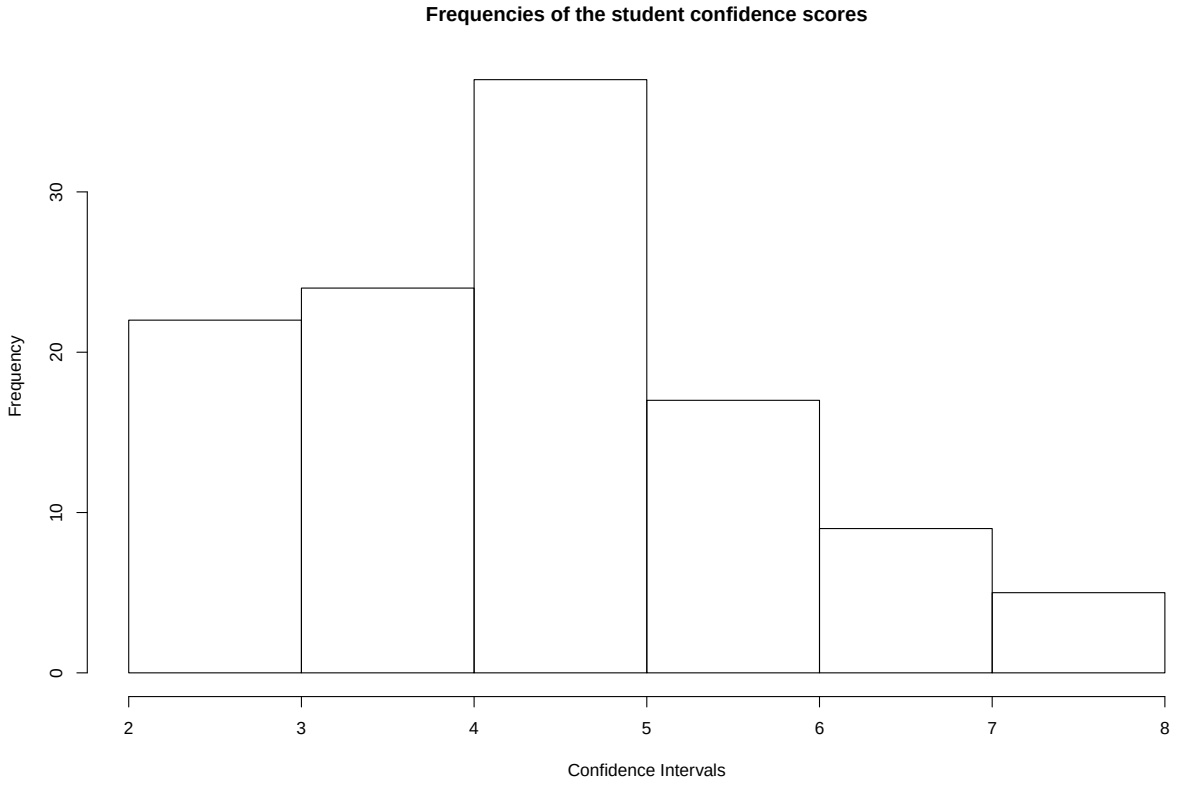
Can the Students be Ranked?

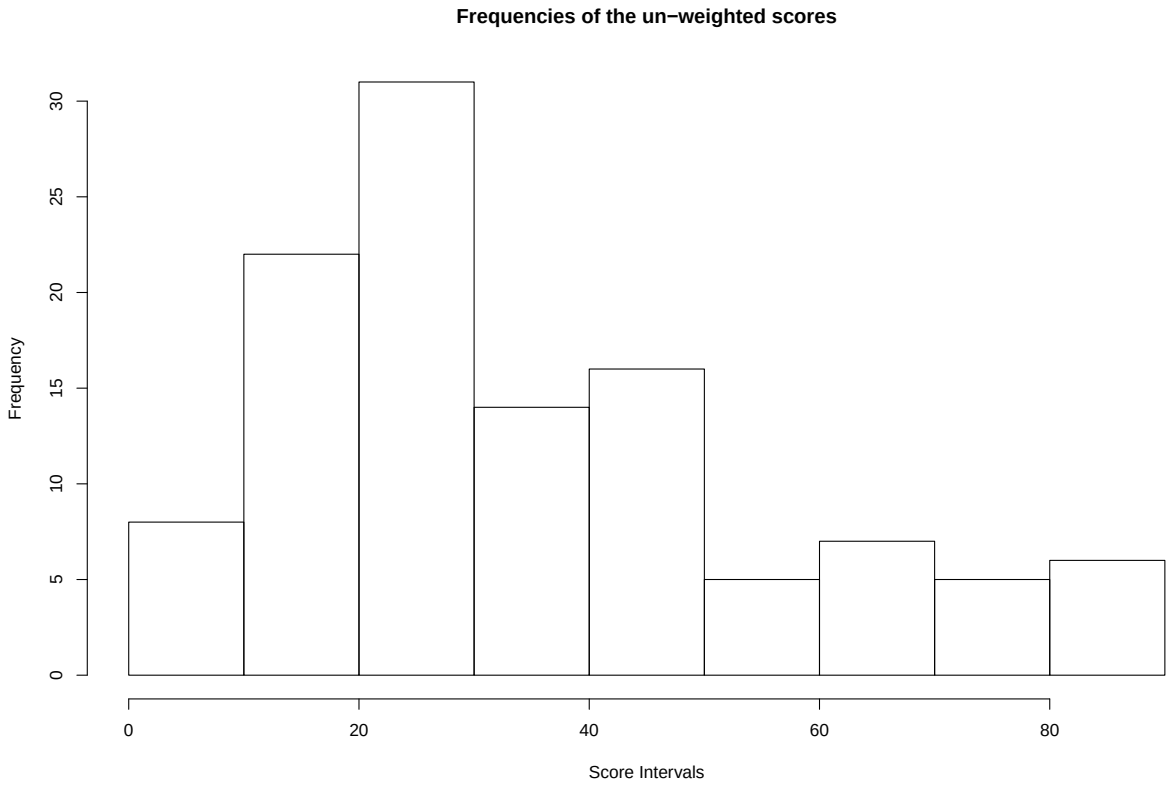
First, we assign a weight to each skill and a confidence score to each student. For each skill (*ggplot2*,*dropbox*,*R etc..*), its weight is inversely proportion to the number of students that have it. For example, *R* has a score of **21** as many students know it while the skill such as *lattice* has a very high score of **97** as very few students know it. Confidence score of the student is his average rating calculated by assigned ratings to each confidence level (Expert = ***4***, Confident = ***3***, A little = ***2***, None = ***1***). We use these two metrics to calculate weighted and unweighted score a student.

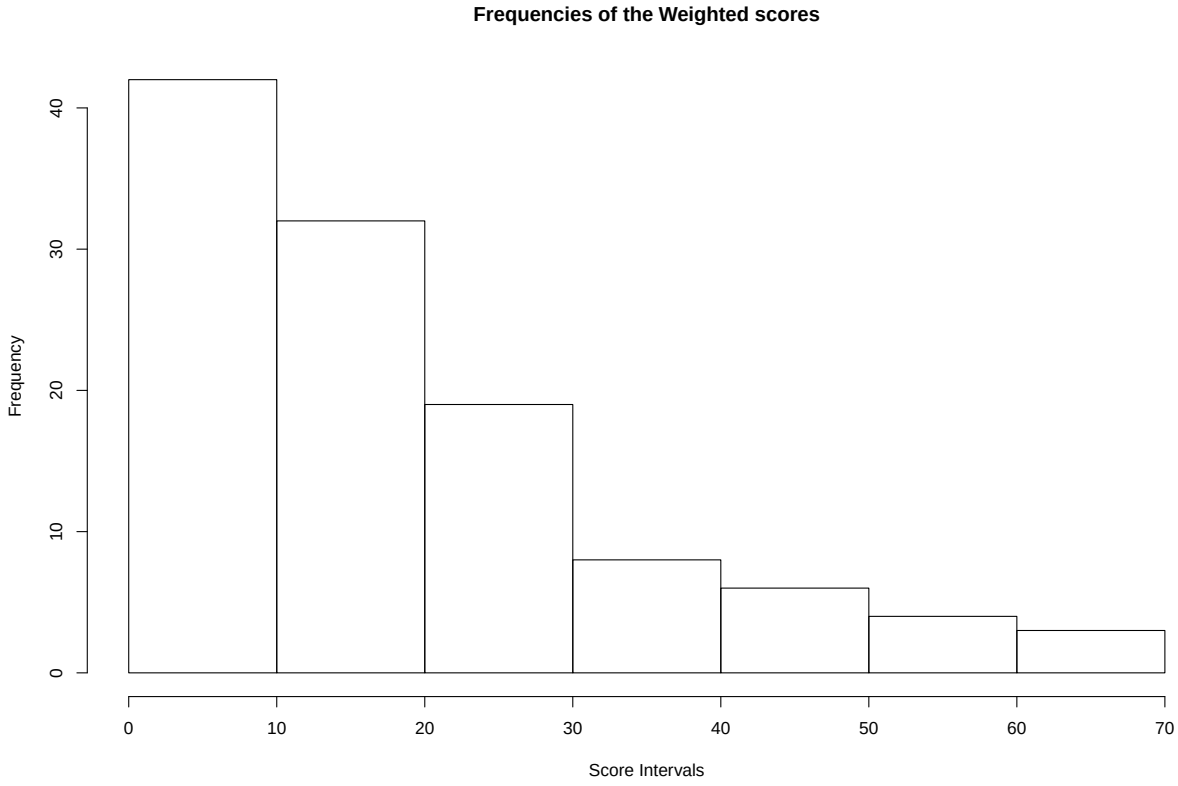
Unweighted score of the student is the sum of all the weights of the skills that he know. Weighted score is the product of unweighted score with the weight of student calculated above using the confidence measures. The graphs below depict the weights and scores calculated so far.

The histograms of the 4 metrics calculated till now are presented below.


Frequencies of the weights of skills
Frequency
0
2
4
6
8
10
0
20
40
60
80
100
Weight Intervals

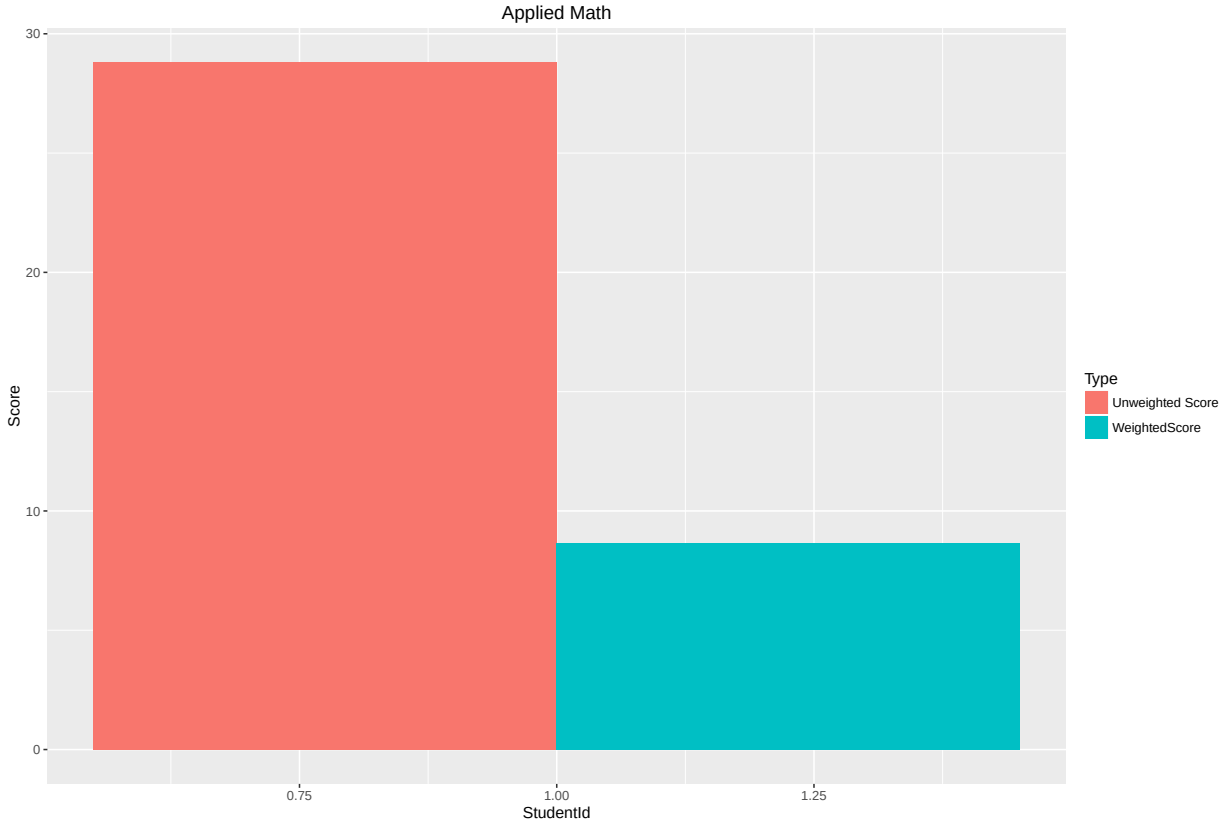

Frequencies of the student confidence scores
Frequency
0
10
20
30
2
3
4
5
6
7
8
Confidence Intervals

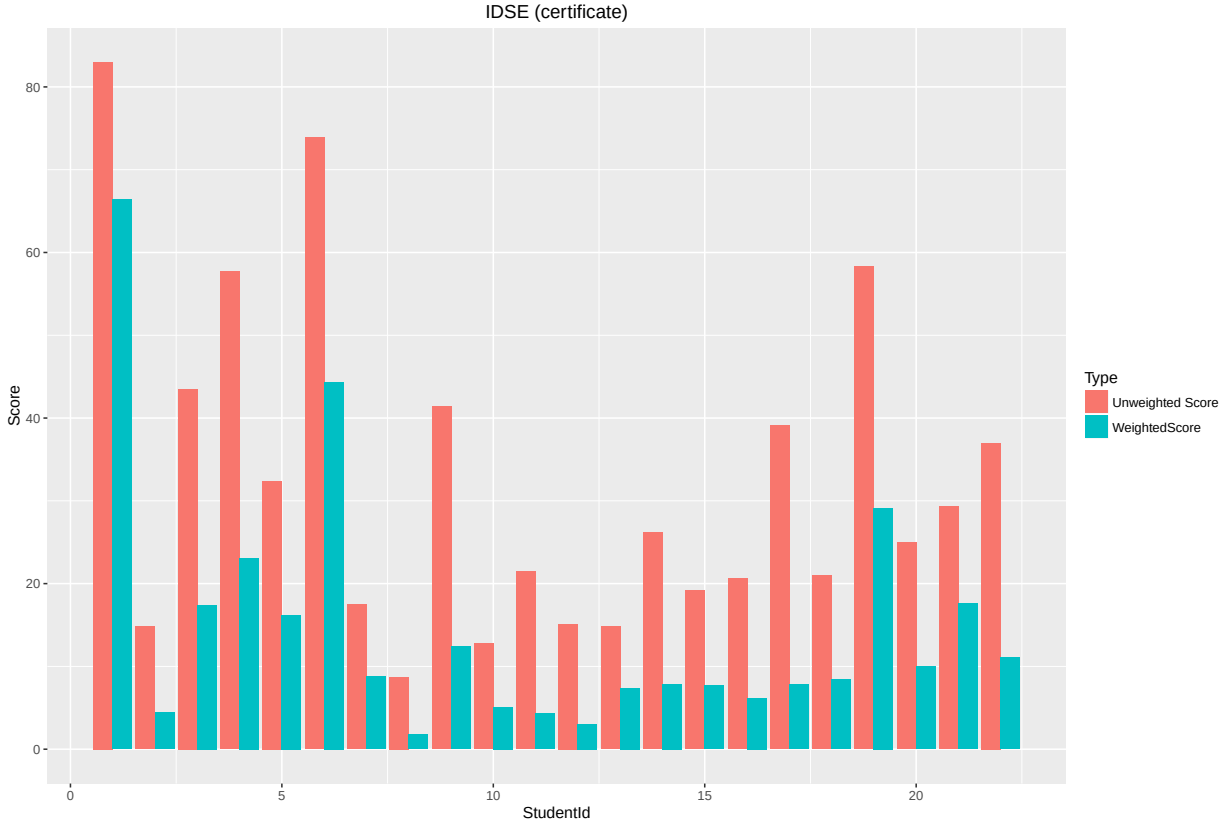




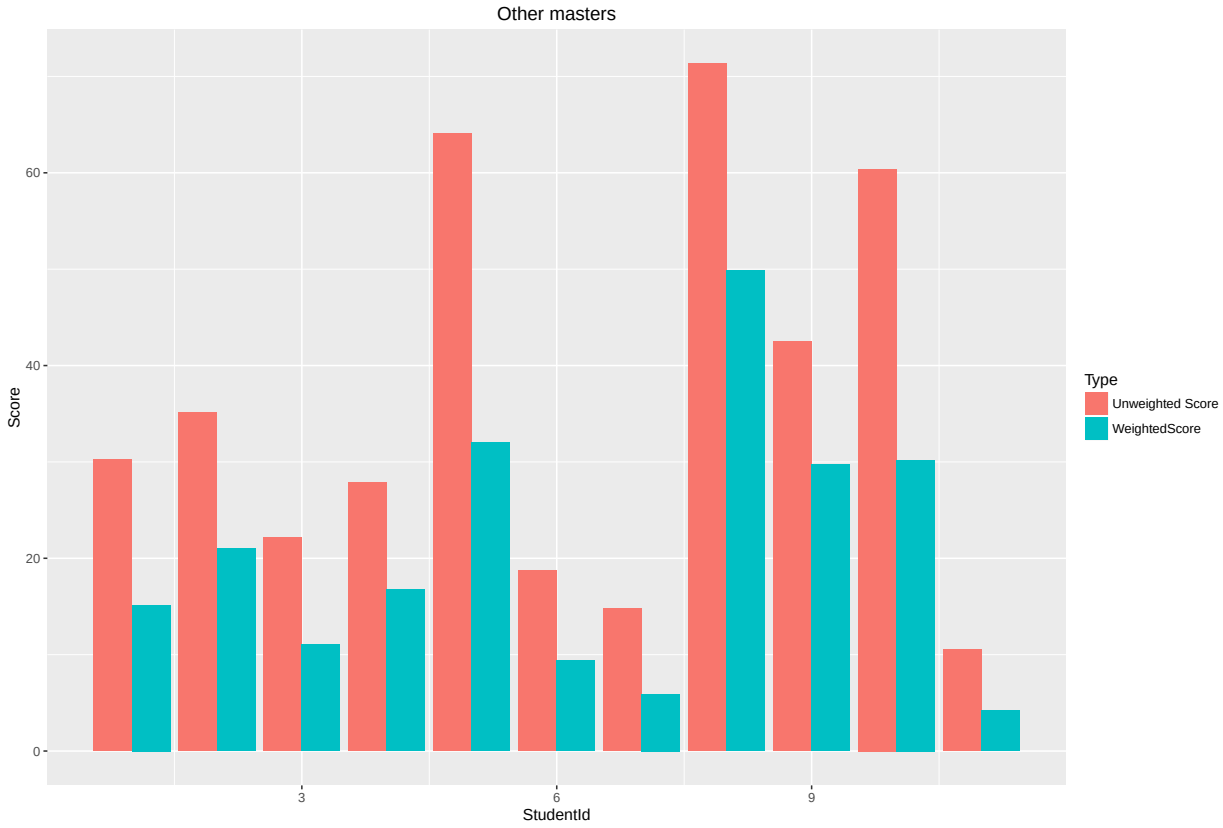
Below, we present the individual score of each student, first segregated on the basis of the program and then

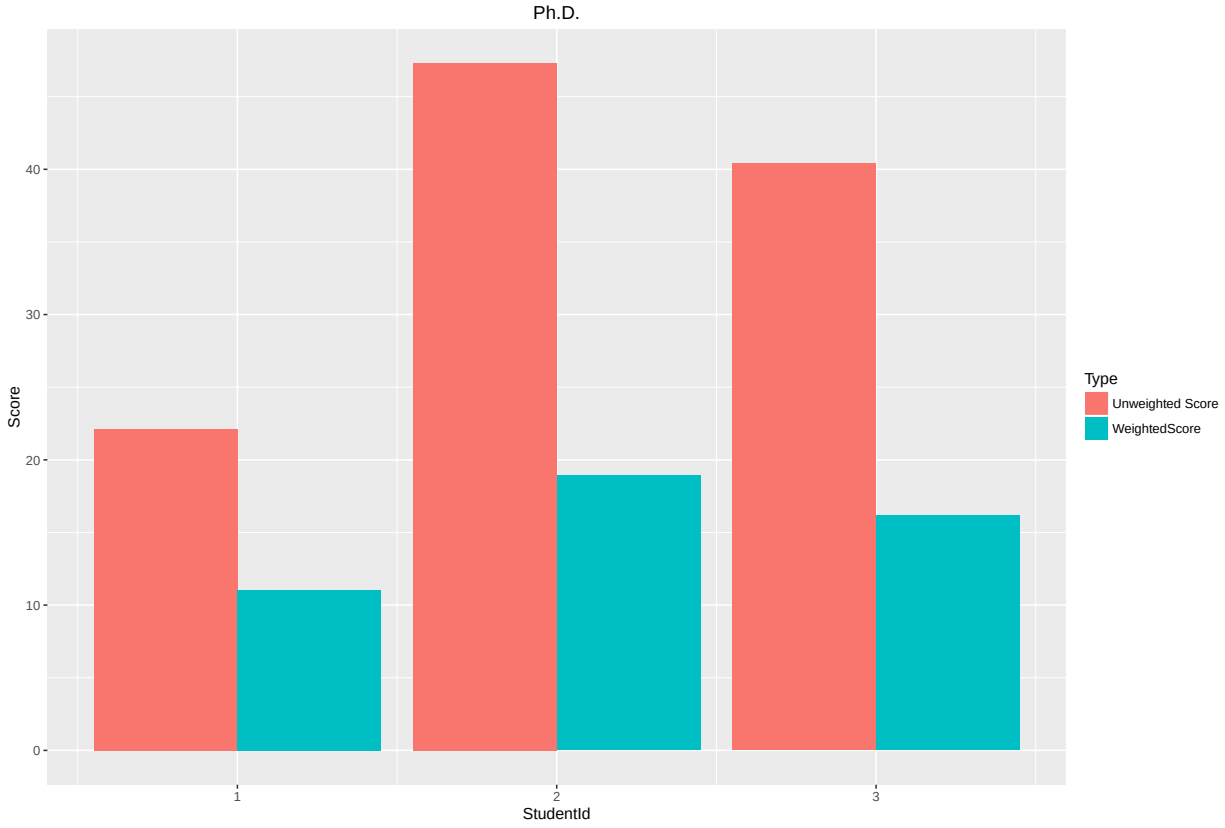
for the entire class.

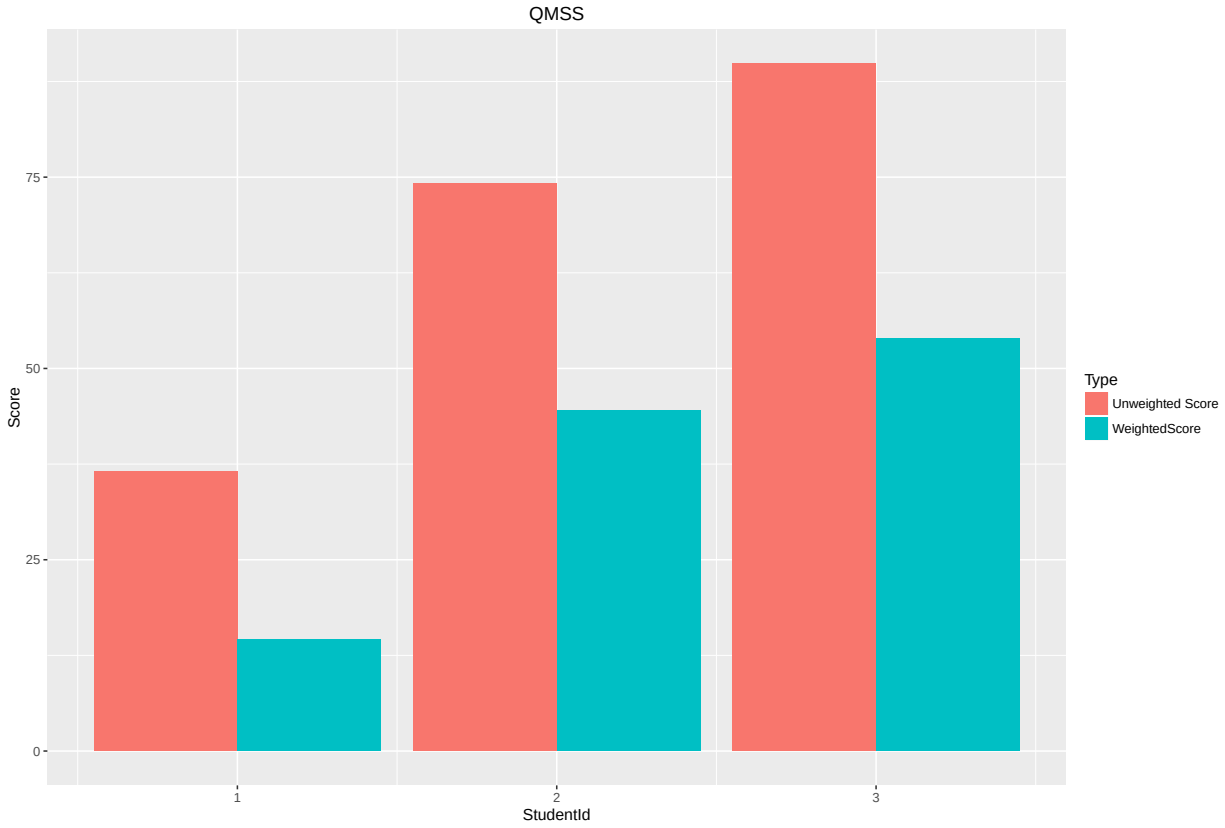


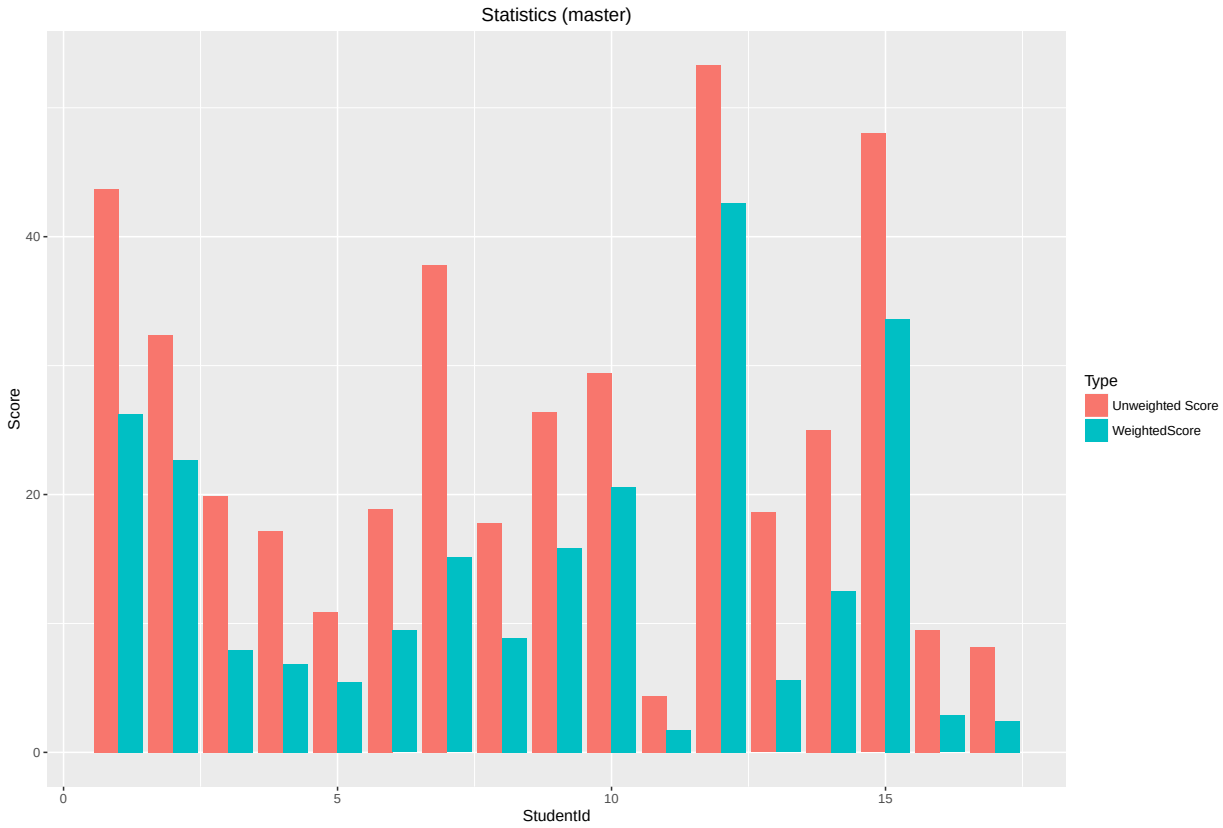


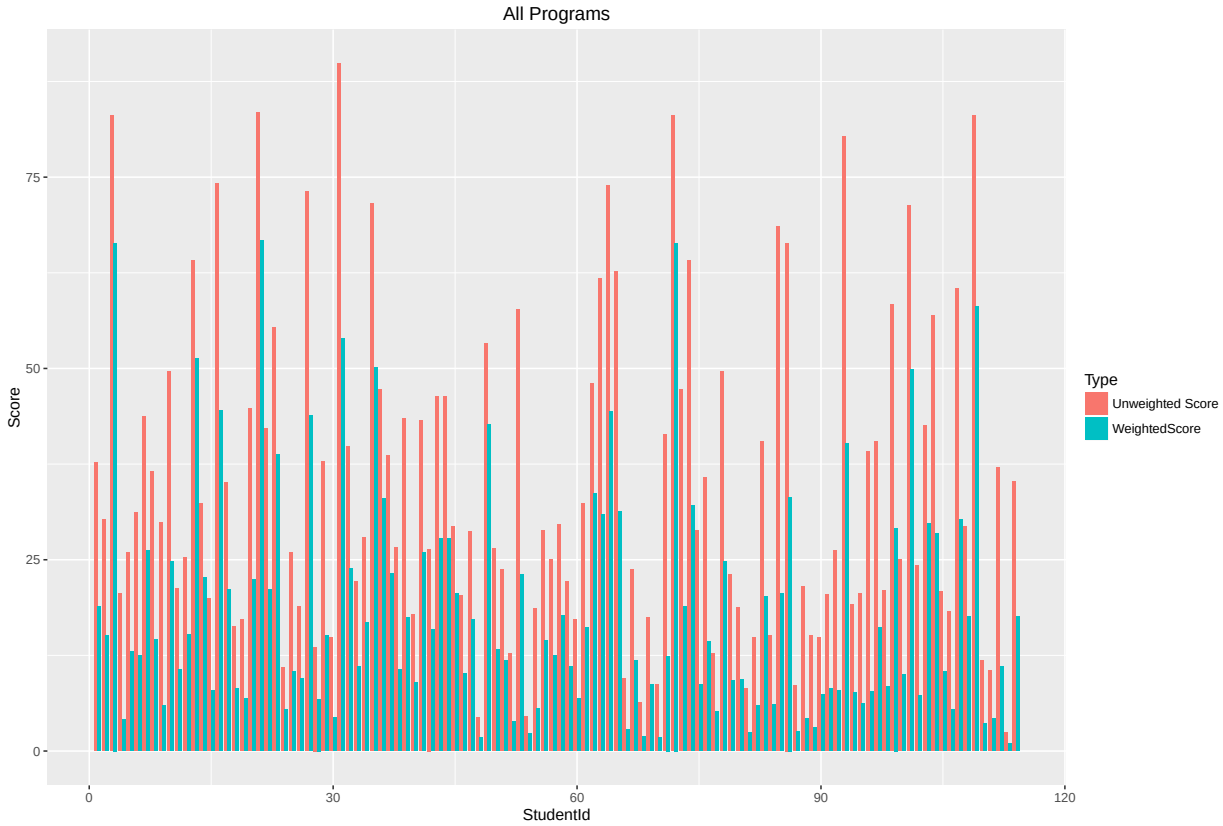

IDSE (master)
Score
StudentId
0
20
40
60
80
Type
Unweighted Score
WeightedScore


Other masters
Score
StudentId
0
20
40
60
3
6
9
Type
Unweighted Score
WeightedScore


Ph.D.
Score
40
30
20
10
0
1
2
3
StudentId
Type
Unweighted Score
WeightedScore


QMSS
Score
75
50
25
0
1
2
3
StudentId
Type
Unweighted Score
WeightedScore





We can observe a trend that most of the students doesn't have a high score, as expected from the diverse

skill set. Also, weighted scores are always less than the unweighted score (which can be inferred from the average lines). In the comparision graph, we can observe that for students who had high unweighted score also had huge decline in the weighted score suggesting that there is a scope for learning even when you have an understanding of the skills. If the course privdes an opportunity for students learn all the skills listed, then the difference between weighted and unweighted scores should be minimized. We can only wait for the survey at the end of the course to ascertain our claim.