EDVA Project I: Course Students Analysis

Manuel Rueda

February 11, 2016

Data Cleaning

Before starting to analyze the data, first we are required to clean it and standarize 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 standarize 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 *Prefered Editor* column contains redudant entries. We proceed to standarize 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
```

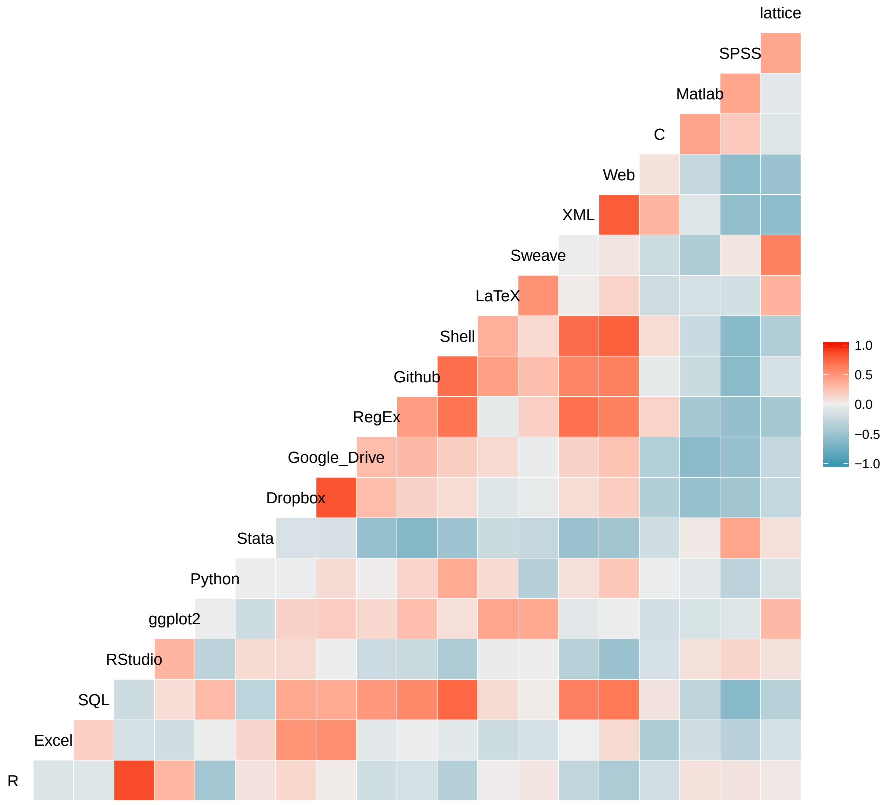
Cluster Analysis and PCA

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.

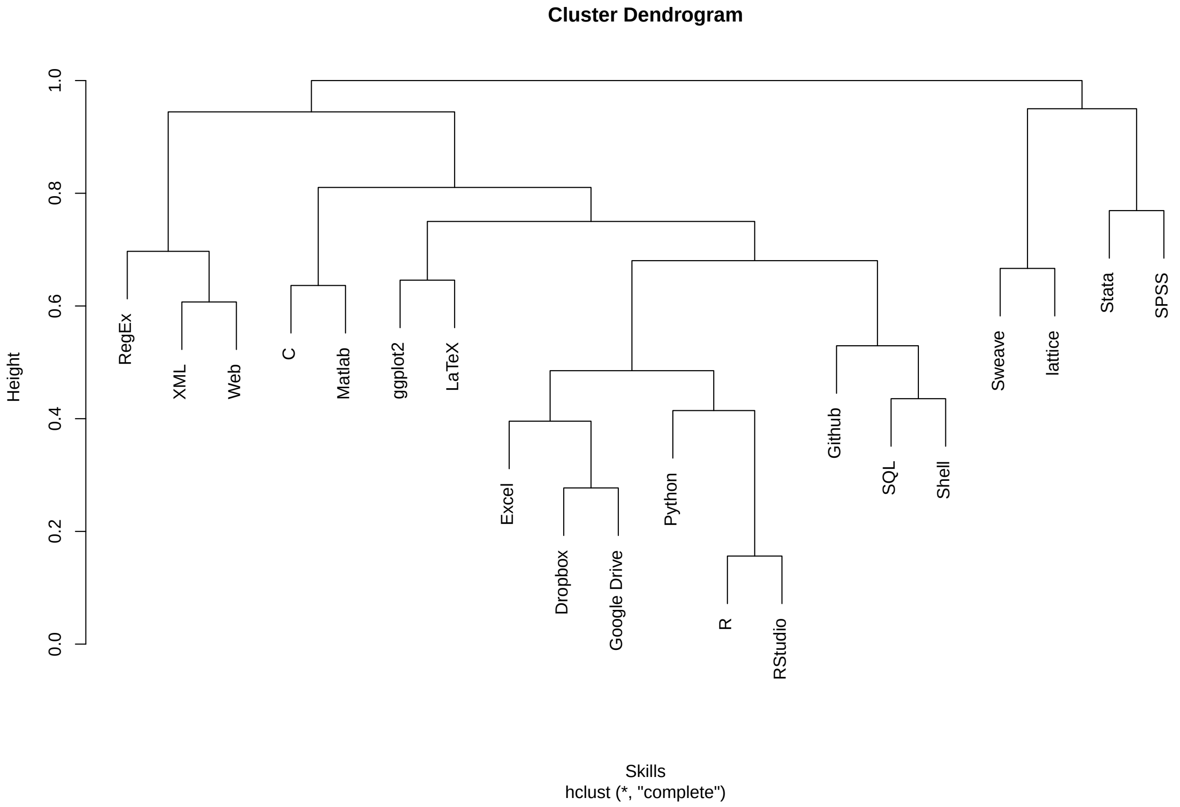
The list of reported skills are:

```
##  [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.



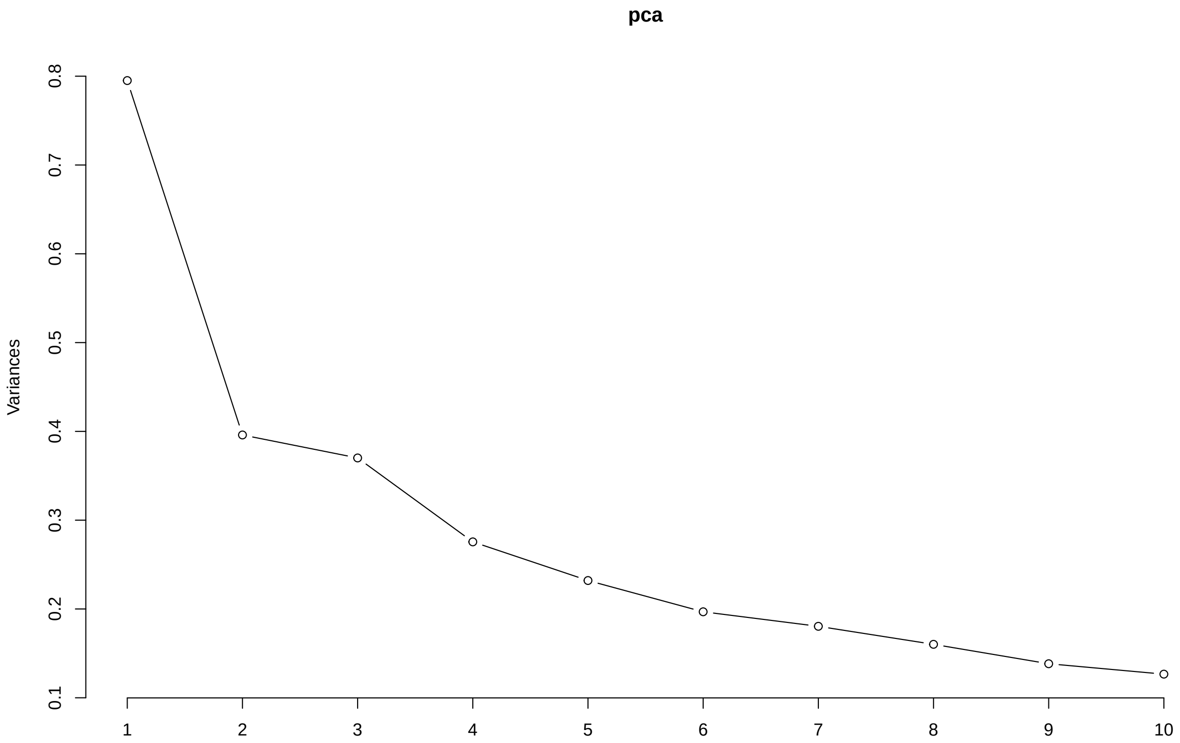
The correlation matrix gives us a good idea of how the skills relate pairwise, but it would be interesting to see how each of them relate to the rest, giving us a broader picture. Building a hierarchical binary dendogram is useful for this.



This picture provides us some interesting insights:

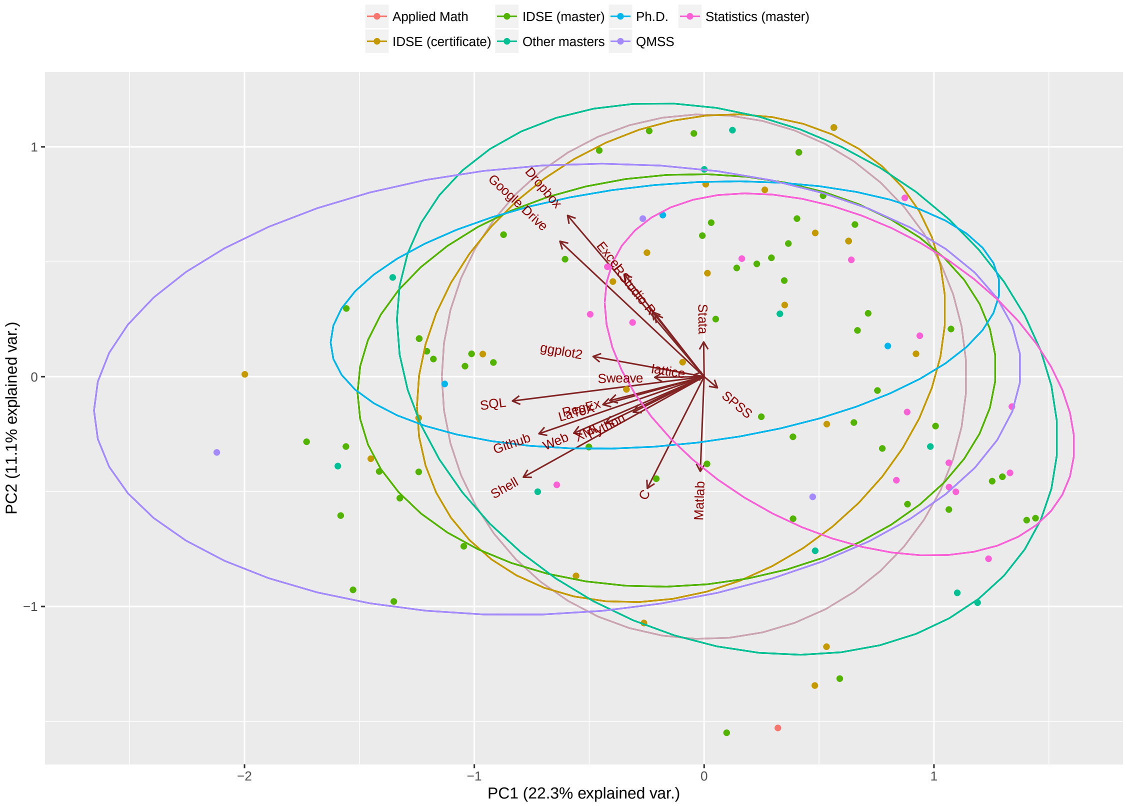
1. As expected, *R* & *RStudio* are closely related, followed by *Python*. These tools can be considered the core of the Data Scientist.
2. *Excel*, *Dropbox* and *Google Drive* also cluster together. These can be considered the least technical of the skills.
3. *XML* and *Web* go together, which makes 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 distirbutions 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, indiciating a dissasociation with them.

Related to the programs,