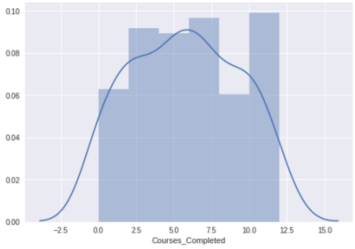
**Introduction: Summary and problem definition for management**

In order for the MSPA program to improve on student experience and stay competitive in the future, it needs to respond to student requirements for skill development as they relate to industry trends. An important element to consider is the integration of data science tools within the curriculum itself. The question for our team then is: which tools are perceived by students to bring the greatest value personally and professionally and how does this influence their enrollment preferences?

This research project is based on a collection of data from the 'MSPA Software Survey', in which the university gathered responses from students and faculty in relation to their own preferences and their perception of industry trends. We hope to be able to draw conclusions from their feedback and use these insights to adapt our curriculum. The research focused primarily on five tools: Python, R, SAS, Java and JS, and trends in student perception of value.

**Discussion of the research design, measurement & statistical methods**

The survey was delivered in 2016 to 207 students who demonstrated a range of experience. The majority of them planned to graduate by 2018, and had an average of 6 courses prior to the survey as demonstrated below.




**Distribution of 'Courses Completed'**

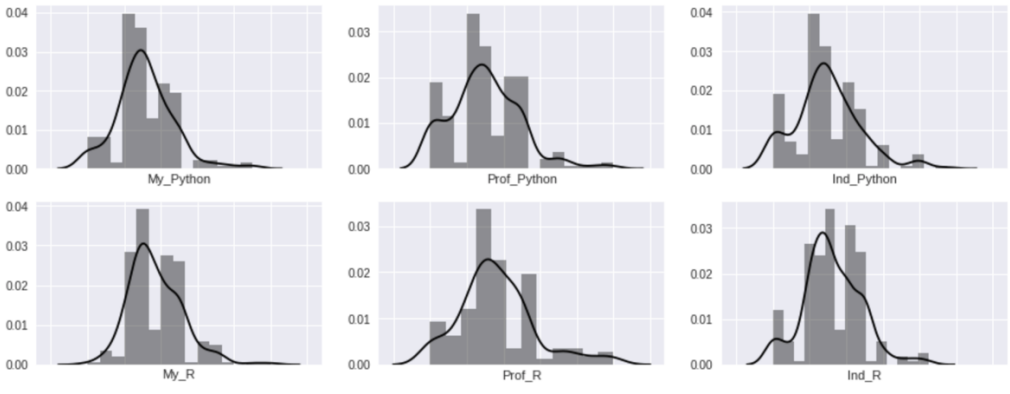
| Exp. Graduation | # students |
|---|---|
| **< 2017** | 85 |
| **2018** | 86 |
| **2019 >** | 33 |

A minority of respondents were just beginning their studies and others had completed as many as the required 12, but at least 70% of the respondents had participated in at least one course that included each of the three core tools deliverd in the program: Python, R, and SAS. The breakdown of course (and relative tooling) can be seen in the table on the right.

| Course_Title | Count | Software | Title | Core? |
|---|---|---|---|---|
| PREDICT401 | 171 | R | Intro to Stats | 1 |
| PREDICT400 | 163 | Python | Math for Modelers | 1 |
| PREDICT410 | 145 | SAS | Regression_Linear | 1 |
| PREDICT420 | 127 | Python | Database_systems | 1 |
| PREDICT411 | 113 | SAS | Regression_nonL | 1 |
| PREDICT413 | 59 | R | Time Series | 0 |
| PREDICT422 | 48 | R | Practical ML | 1 |
| PREDICT455 | 30 | R | Data_visualization | 0 |
| Other | 26 | NaN | Other | 0 |
| PREDICT450 | 17 | R | Mktg_Analytics | 0 |
| OtherR | 14 | R | Other R | 0 |
| PREDICT452 | 13 | Python | Web_Analytics | 0 |
| PREDICT453 | 11 | Python | Text_Analytics | 0 |
| PREDICT451 | 7 | R | Risk Analytics | 0 |
| PREDICT456 | 6 | R | Sports_performance | 0 |
| PREDICT454 | 5 | R | Adv_modelling | 0 |
| OtherPython | 5 | Python | Other Python | 0 |
| PREDICT457 | 4 | R | Sports_mngmt | 0 |
| OtherSAS | 2 | SAS | Other SAS | 0 |

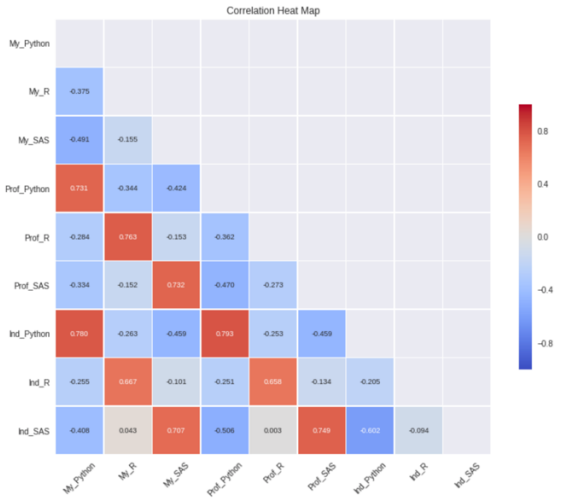
Within the survey students were asked a variety of questions. First, they were asked to distribute 100 points between the 5 tooling options: Python, R, SAS, Java and JS in relation to three different questions – 1) their personal interest in the topic, 2) their perspective on the required skill in their profession and 3) their experience in having to use the tool in their industry. Following these questions, students were asked about their interest in certain curriculum topics: Python, Foundations in DE, Analytics Apps, and Systems Analysis.

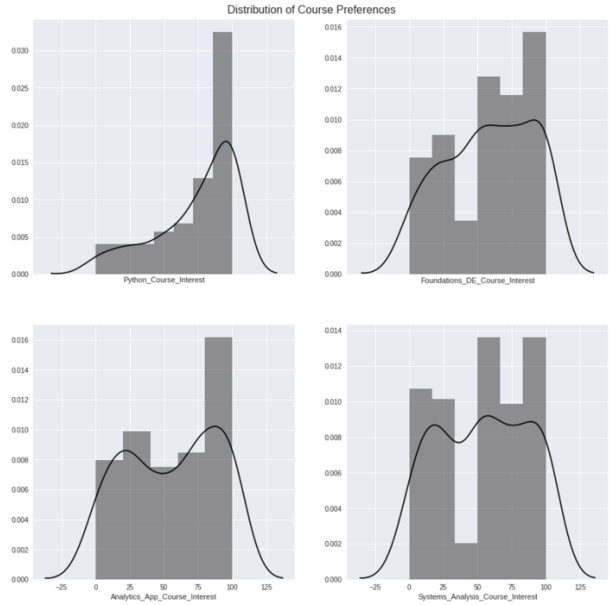
To begin, the exercise in which students distribute 100 points across the 5 tools in three unique topics demonstrated that the emphasis from a student perspective was on Python and R over SAS, Java and JS. Relative to the other distribution plots, these showed a clear skew towards normal, and away from a lopsided histogram towards the 0 measure on the left.




**Distribution plots of the student Reponses for Python & R: personally, professionally and in their industry**

The resulting correlation heat map between the student responses show the highest correlations in the fields of the 3 core tools, with the most correlated fields showing as warm colors. Java and JS did not show at significant levels outside those respondents who had a niche interest in these tools.



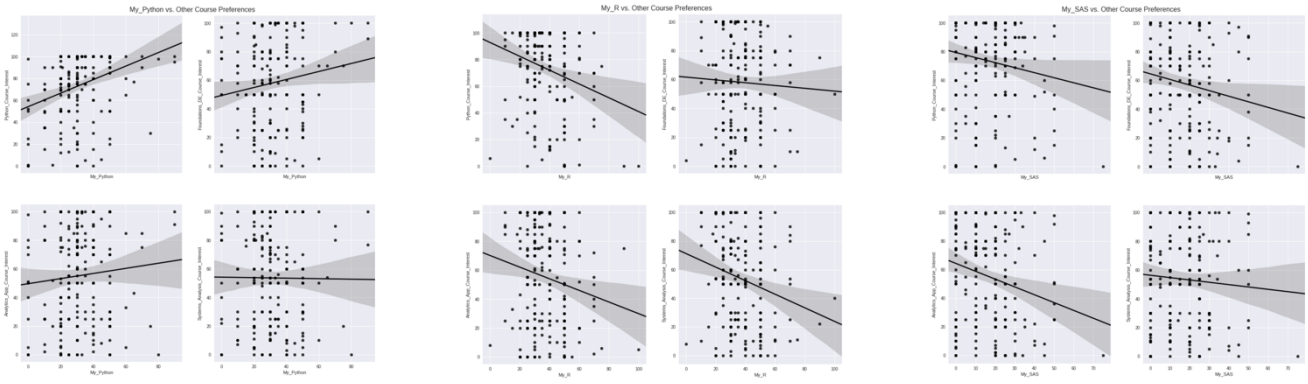

In addition these distribution plots above demonstrate the range of 'course interest', which also skews in favor of Python content - the top left shows that over 30% of respondents have a strong interest in Python courses, while other focus areas top out at 15%.

**Overview of the programming work**

The Python notebook that was created for exploratory data analysis and initial regression can be found in the Google Colab file linked below.[1] Raw data from the csv file was loaded into the notebook and then the analysis was divided into three main areas: first, the student responses to the requirement to distribute 'points' across the five tools in three topics. Second, the student expressed interest across the four main curriculum areas (Python, App Development, Systems, and Data Engineering). Third, initial regression views of correlation between variables, such as student preferences relative to perceived industry trends.

The focus on the programming work was to explore the relative data fields, correct any missing values, understand the differences in scale and then investigate relationships within the data. As shown below, early regression plots demonstrate that the strongest positive correlation is with the 'My_Python' responses against future course opportunities in python or DE (across the top). As the discussion becomes about app development or systems analysis (across the bottom) there is less interest.

**My_Python** **My_R** **My_SAS**



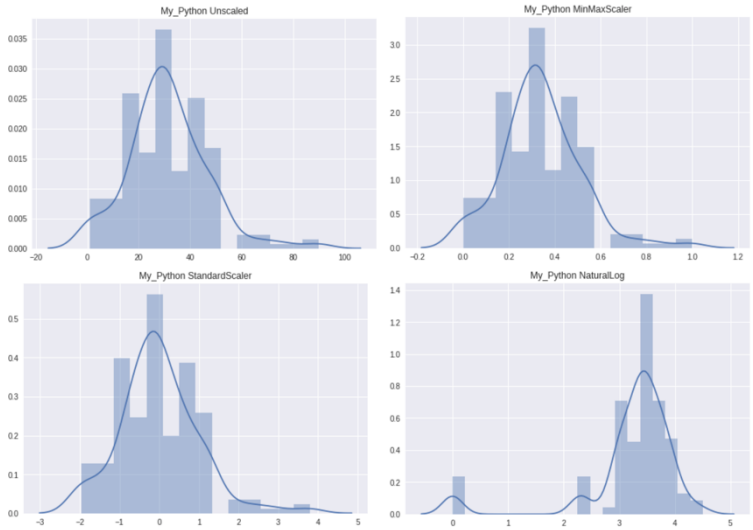

**Stated interest in program tooling relative to curriculum preferences**

[1] Google Colab Notebook: https://colab.research.google.com/drive/1NcEcIRRoJr5OViEFCDcQ-E9rQ1WNliJH

It should be noted that the only positive correlations are with the 'My_Python' group. All other tools show a negative trend in relation to possible curriculum options, which may be an indication that we need more to offer these students.

Finally, within the notebook is a review of the effect of transformation against the variable 'My_Python', which was the variable with the strongest correlation values. Using the Seaborn.distplot function, we are able to see the effect on the distribution of the data when we use different scaling techniques against any of our variables. For the most part, the distribution seems un-effected when using the 'Standard' or 'Min-Max' scalers, but then the distribution becomes much more skewed to the right when using a natural log transformation. In fact, to use this transformation we needed to update any zero values in the data to a small value to be able to run the transformation. This resulted in a much different silhouette to our distribution plot. The variation is shown below:

**Distplot of 'My_Python' – Unscaled and then using 3 tranformation methods.**




**Review of results with recommendations**

The initial review of the data suggests that there is a strong preference for using Python as the primary data science tool. This appears to be a student preference, but it is also supported by the evidence that students perceive a strong demand for these skills in the market. As we consider the current applied tooling that is emphasized by the MSPA curriculum, there is an equal balance across Python, R, and SAS – with the elective courses such as Marketing (450), Sports Analytics (456/457), Data Visualization (455), and Time Series Forecasting (413) placing the 'specialty' emphasis on R as the tool of choice.

Although it is good to have a balance, we may consider more ways to bring Python into the curriculum. Judging by the fact that this course (Practical Machine Learning – 422) is now taught in Python may be an indication that the program is moving in that direction already.

**Appendix :**

Included in the following pages is the pdf version of the python notebook used to create these images and the full pdf versions of the graphics found above.