Introduction

In December of 2016 the MSPA student population was surveyed to assess interest in specific programming languages and in potential new class offerings. The goal of the survey and associated data analysis is to inform curriculum decisions regarding programming languages and course development for the MSPA program. Free-form text data was also collected by the survey instrument, but is not part of this analysis.

Methodology

The survey was conducted via an online survey tool, and the responses form a convenience sample of students who elected to participate. Questions were asked about the students' interest in a range of programming languages, professional need for those languages, and the importance of those languages in their workplace. Students were assigned a 100 point "budget" for each question, and asked to allocate the points across the options to indicate relative importance; this yielded integer values of relative interest. Checks built into the survey tool require all 100 points be allocated for a question's response to be valid and allow submission, ensuring a consistent scale on all responses. This "point budget" assignment method was also used to gauge interest in 4 potential new classes.

Responses were analyzed using several different techniques. Simple summary statistics were generated for each of the languages surveyed, as well as for the number of courses respondents had completed in the MSPA program. Data were visualized using bar plots, heat maps, and boxplots. Finally, data were explored via Clustering, Principle Component Analysis (PCA) and OLS regression techniques.

Programming

The analysis is coded in Python 3.6, using standard packages available through normal Python distribution channels. The code has four main sections. The first explores language interest in various ways. Next, PCA is used to look for potential multicollinearity issues. This is followed by OLS regression to see if there are predictors for class interest, and finally clustering to see if that offers insight.

Results and Recommendations

As seen in the R-vs-Python scatterplot[1] most students have mixed preferences relative to R and Python, with a few showing strong preference for just one of the two. Aggregating the language variations into families[2] shows the R family has the highest mean preference, closely followed by the Python family; SAS, java and HTML/CSS lag Python by a significant amount. A new class in Python has the highest mean score; all 4 proposed classes have a mean interest >50%[3]. There is high correlation between interest in Analytics App and System Analysis classes; as well as between Analytics Apps and Foundations.[4] PCA showed no signs of multicollinearity; so, the data set is not deficient.

The analysis failed to find a predictor (or set of predictors) for class interest or language preference via OLS regression[5]. The individual language families were tried as response variables, as well as each of the proposed classes. The best OLS models accounted for less than half of the variation in the data set, which makes them questionable for basing curriculum choices on. Clustering revealed a relationship between web-coding language affinity and interest Analytics App, System Analysis and Foundations classes. People with affinity for Python clustered to interest in Python coding classes[5]. R and Java/Scala adherents do not cluster to having any particular class interest.

Continuing to offer both R and Python as languages in the curriculum is supported by the data. All 4 proposed classes show interest from a significant number of respondents as well. The recommendation is to focus on R and Python offerings, and go ahead with the proposed new classes.

[1] See file plot-scatter-r-python.pdf
[2] See file barplot_agg_sw_interest.pdf
[3] See file boxplot_class_interest.pdf
[4] See file plot_corr_lang_class.pdf
[5] See file Console_output1.html