Assignment 1 - Exploring and Visualizing data

Executive Summary & Problem Definition

In order to keep up with the rapidly evolving data science landscape, a survey was conducted among Northwestern graduate data science program students. The focus of the survey was to understand the programming languages most relevant to data science industry and students desire to learn. In addition Northwestern was seeking to understand interest level in offering new electives to expand the program.

Results from this survey will help decision making for the graduate program around future course offerings that are relevant and meet student's interests and identify potential areas of improvement in course material, with primary focus on understanding programming language interests.

Research Design

The survey was split up into four main sections: individual preferences for the main programming languages offered today (personal desire, professional needs, and where students see the industry today), exploration of new course interest to the program, open ended response of improvements to the curriculum (not part of the analysis and scope of this assignment), and respondents' current progress within the graduate program.

Our analysis focuses on several key areas to help determine the future direction of the graduate program:

1. Preferences and trends of courses and programming languages
2. Appropriate focus and allocation of curriculum and content of the graduate program to these trends
3. Potential interest of prospective students and new courses to the program

To address these key areas, we will analyze survey questions among total survey respondents and also look for patterns across segments of respondents such as based on "courses completed", "anticipated graduation date".

Technical Overview

Python is one of the most prominent programming languages in the data science industry and is the language used for this analysis. The analysis uses popular python packages such as pandas, matplotlib, seaborn, and numpy to help analyze and visualize results.

For this assignment, the dataset is relatively small so traditional statistical methods can be employed via python to analyze results. The use of only one or two key variables at a time is needed where recommendations can come very quickly from classifying and grouping student responses into a few key variables for identifying patterns and trends. Transformation and scaling of the data may also be needed for discussion and presentation of findings to leadership for review.

Key Trends

R programming language is by far the most preferred language among survey respondents. Python follows R as the second most preferred programming language. SAS is third, but at large step down in the mean and median compared to Python and R. Java and JavaScript are much further down in terms of preference, but there are a few outliers that rated these languages as important where specialization in these areas exist.

There are no significant differences across preference for a specific language by industry/personal/professional interest. When students indicate a preference for a certain language, there is a strong correlation of interest across personal, professional and industry preferences.

We do not notice a strong correlation between course completion and preference for programming languages which indicates that student preferences are independent of course progression.

Conclusion and Recommendations

The following are recommendations based on our data analysis:

The new classes offered should reflect both the students' preferences as well as current and prospective professional needs. By the students' input, R & Python should be the primary languages used to teach classes, and SAS can be another option. However, SAS should be less common. The program can oscillate offerings between what language to use for which class in the following quarters and monitor both student feedback and enrollment data. The primary languages will be R and Python, but some terms can also input SAS sparingly.

It's recommended that faculty and staff responsible for maintaining the MSDS curriculum take into consideration the findings of the survey and overhaul the program to keep it relevant to the professional world today.