Assignment 1: Exploring and Visualizing Data by Mimi Trinh

Section 1: Summary and Project Definition

In order to inform the data science curriculum planning, an online MSPA Software Survey was launched in December 2016 using Survey Monkey. With the objective to define the future direction of the graduate program, the survey was developed with the following goals 1) learn about student software preferences and their interest in potential new courses 2) guide software and systems planning for current and future courses.

Section 2: Research Design, Measurement, and Methodology

The project is divided into two main parts 1) data visualization and exploration 2) data transformation and scaling. First, the raw data in csv format was loaded into Python for data cleaning, preparation, and visualization purposes. From examining the dataset, there are 207 rows or responses and 40 columns, which represent the respondent ID and survey questions. We dropped the respondent ID column as a variable and instead used it as the index of the data frame to label each row. Multiple scatterplots were created to examine the bivariate relationship between each pair of software preference i.e.: personal reference for R vs. personal reference for Python. Then a correlation heatmap was created to visualize the strength and direction of correlation for each pair. Next descriptive statistics was presented for each software preference, courses completed, and courses interested along with visualizations for each variable. Finally, the variable of Courses Completed was chosen to examine the effects of three transformations (scaling methods): standard, min max, and natural log.

Section 3: Programming Work

The project utilizes multiple packages in Python: pandas for data frame operations, numpy for arrays and math functions, matplotlib for static plotting, and seaborn for pretty

plotting, including heatmap. The original starter code was provided in the text file py format, and the working code was conducted in the ipynb Jupyter Notebook format. Some of the column names were shorten, including the software preference variables i.e.: changed “Personal_Python” to “My_Python.”

Section 4: Results and Recommendations

First, the heatpmap shows some interesting correlations such as a positive relationship (0.763) between personal preference for R and professional preference for R. It shows that personal, professional, and industry preferences for each software move together.

Second, the descriptive statistics show that the variable of Courses Completed has the mean of 6.34, which shows that the expected survey respondent is half way through the MSPA program. Also, using the mean of each software preference in descriptive statistics, the order of most to least preference is R, Python, SAS, Java, and JavaScript. As a result, it’s recommended that the data science curriculum incorporates R and Python, and possibly SAS, while it’s safe to drop Java and JavaScript from the program. Using both the mean and median, the order of software preference ranking remains the same, so we can conclude that there may be some skewness in the dataset, but the skewness is not serious.

The descriptive statistics show that courses with the highest mean are Python (73.53), following by Data Engineering (58.05). So it’s recommended that management should provide these options as future classes for students in the program. Finally, three transformation methods were applied on the Courses Completed variable. Standard and min max scalers didn’t change the shape of the distribution, but natural log did. It’s best to use scaling methods that preserve the shape of the distribution, so the first two methods are recommended over natural log.