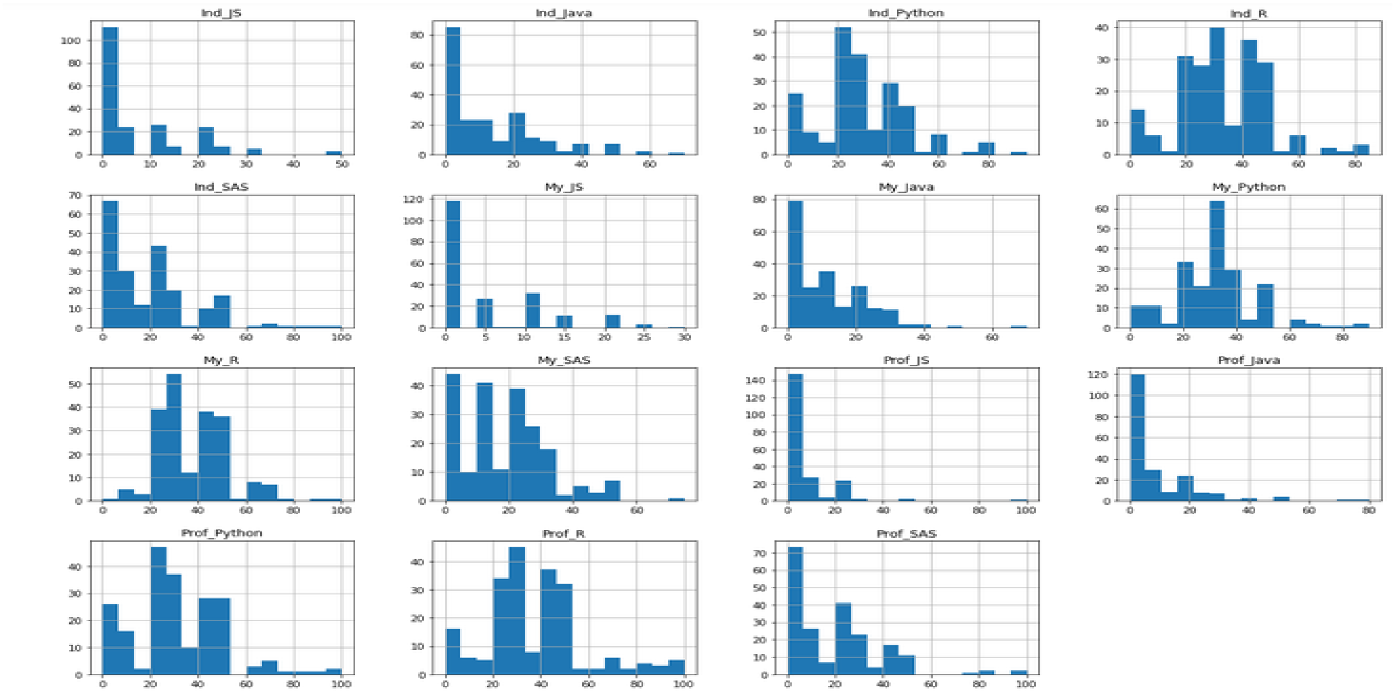
MSDS 422 Assignment 1

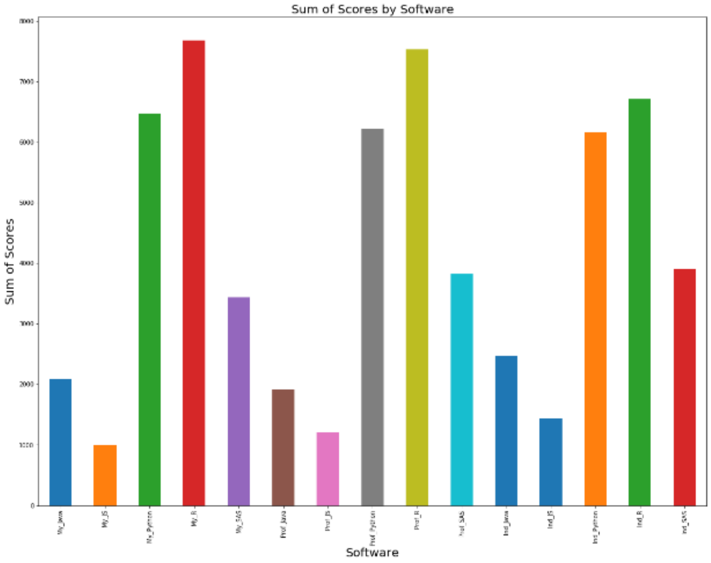
The MSDS software survey is a useful tool for academic administrators to guide the course curriculum for future direction of the graduate program. The survey asks students to allocate 100 points to determine which software (Java, JS, Python, R, and SAS) they prefer for personal, professional, and industrial use. It then goes on to ask students to score potential classes that they have interest in. By performing some statistical tests and visualizing the data, one can get an idea of which software is preferred and which class has the most interest.

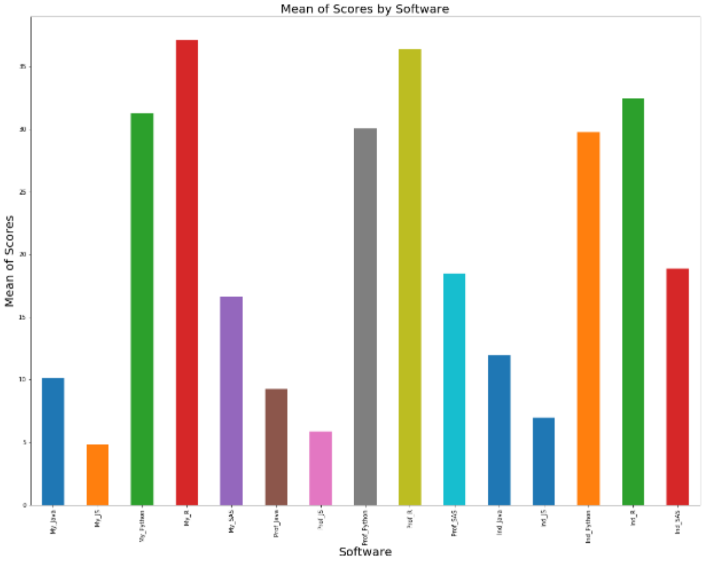
Some preliminary statistics and graphs were made in order to get a better understanding of the data, and who responded. A total of 207 students filled out the survey in 2016 and a majority of students were reported to graduate in 2018 or later. Furthermore, the average student completed about 6 classes. When observing the distribution of interest in personal Python amongst the amount of courses completed, it appears to be fairly well distributed amongst all students. When using transformation techniques to account for outliers and scaling inconsistencies, the minmax and standard scalers have no effect on the profile of the distribution, in contrast to the natural log scaler which alters the shape.

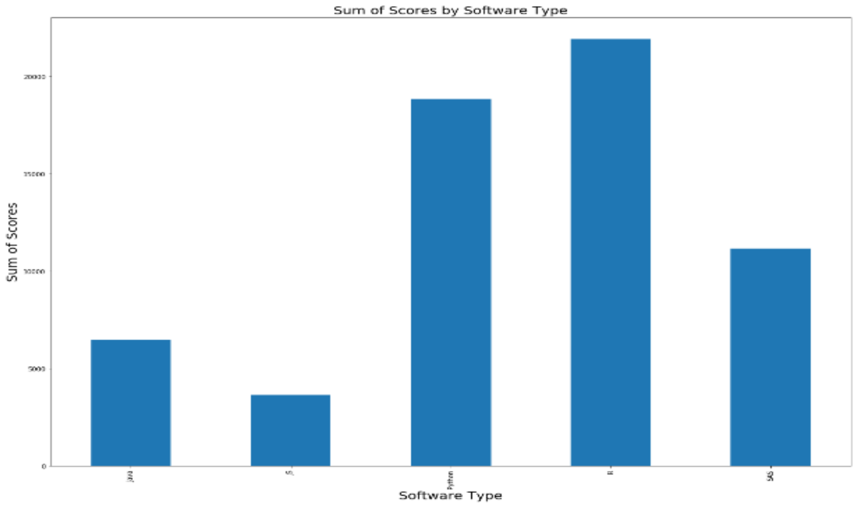
In terms of software preference, it appears that R and Python are the most preferred amongst all categories. This can be seen by observing descriptive statistics, bar plots, and correlations. Amongst all categories, R was ranked highest for personal, professional, and industrial uses, followed by Python. Both the sum of scores and means of scores were highest for both of these software, as seen in the following graphs. This first series of histograms show a distribution of the scores given to different software. One can recognize that R and Python have the most scores in the 20 – 50 region, while other software choices tend to most of their scores in the 0 – 20 region.



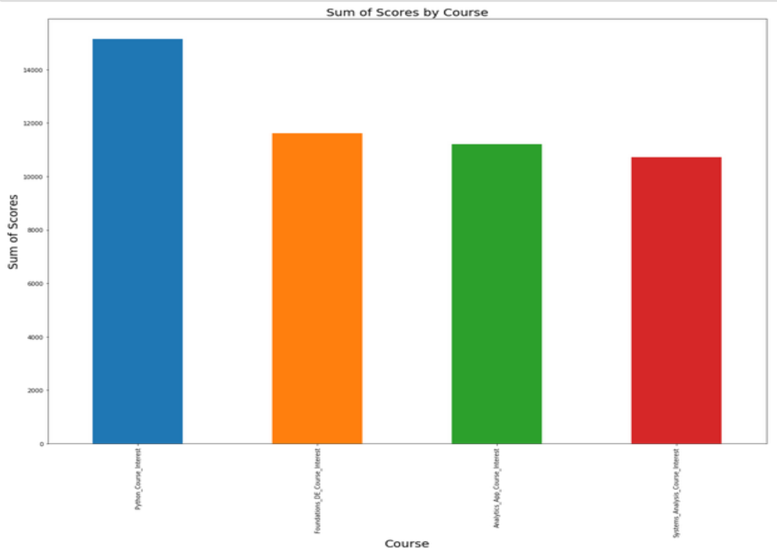
The next three plots display the sum and mean of scores amongst the different software choices. Similarly, R, followed by Python, has the highest rating amongst all categories with the largest totals and largest average scores. Lastly, if one combines the scores of personal, professional, and industrial usage, one will notice that R and Python has the highest combined total.

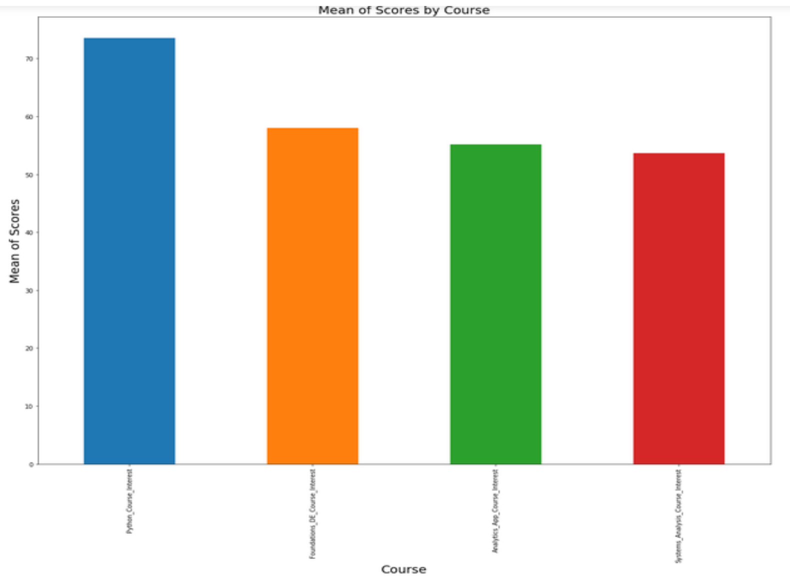






When looking into which of the new courses are preferred, the Python course has the most interest. The other three courses' scores are very close compared to each other. The Python course has the highest mean score and sum of scores, and it is trailed by the other three courses.





When it comes to determining which software is preferred and the amount of interest in the four new classes, one can make the following conclusions. R and Python are definitely the most preferred software, and the Python class has the most interest. Therefore, when moving forward, it may be a good idea to implement more R and Python focused courses. Although the curriculum administrators should have a variety software that is used, more Python classes should be added since it is becoming the preferred language of choice amongst programmers and practitioners.

The Python class should definitely be added to the curriculum due to the amount of interest it received. However, the other three courses also received enough interest for these classes to be added as well.

PDF file of code and output and Python workbook can be found in additional files in folder submitted.