Ypark79 Module 3 Writing Assignment: Limitations.pdf

There are several inherent limitations when scraping and analyzing data from an open source containing anonymously submitted data. These limitations stem from two overarching issues: 1) the lack of standardizing the actual data submitted and 2) the lack of confirming the validity of the data. Examples of the first issue include varying formats of the data submitted (i.e. date formats, scores, GPAs, spacing), misspellings, incomplete information, and varying ways to spell strings (i.e. acronyms or fully writing out names of colleges and programs). This creates the potential of incomplete scrapes and ultimately skews analysis in the long run since the data pool does not contain a complete pull of all metrics.

The second overarching issue of the inability to validate anonymously submitted data is tied to why I was surprised by the outputs of the analytics. Since there is no way to confirm the validity of anonymously submitted data, it is very possible (and likely) that students will report inflated/exaggerated data since it is viewable on open source and potentially due to "self-reporting bias." It is also possible that students only share their successful metrics while omitting their unfavorable ones, which will skew analytics on the publicly contributed data. In a similar vein, the metrics is likely further inflated because people with good scores are more likely to submit their data than those with bad metrics. Thus, it does makes sense that official testing statistics will be lower than the analytics conducted on the anonymously submitted data entries. This means that standardized data sets provided by official testing organizations will be far more accurate than scraped data sets from open source forums like gradcafe. However, the benefit of utilizing datasets from open source forums is that they are easily accessible, free, and can still provide sizeable and valid datasets to conduct analysis to generate valid trends. This is assuming the robot.txt permits the data scraping.

In conclusion, while scraping and analyzing open-source datasets has its limitations, it is still an effective and worthwhile endeavor for data analysis.