Module 3 Assignment - Limitations

Name: Helen Zhang (hzhan308)

Questions:

Having carried out this assignment, please write two paragraphs about the inherent limitations of carrying out analytics over anonymously submitted data items. Did the analytic responses surprise you? How does this different from standards? For example, the average GRE quantitative reasoning score was 157 for 2023-2023 and was nearly 165 for grad school entries submitted (see sample output). Why do you think that is? What might cause this to occur? Please place your essay into a file called limitations.pdf

Response:

There are a lot of limitations to running (and especially presenting) analytics over anonymously submitted data items like GradCafe. With self-reported grad admissions, my database will never represent all applications or outcomes — it only reflects what people choose to post on the site, which isn't the same thing as true admissions trends. A big issue is selection bias in who posts. For example, people who are admitted may be more motivated to share their results, while people who are rejected may be less likely to post (or may post less detail). Also, applicants to more high-profile/popular schools are probably overrepresented compared to smaller or lesser-known programs. So if accepted students post more frequently, the acceptance rate I calculate from GradCafe entries will likely be higher than the true acceptance rate.

Some of my analytic results surprised me. I expected more applicants in my database who applied to Computer Science at JHU and other well-known schools, but the counts were lower than I thought. The GRE and GPA averages didn't surprise me as much, since I assumed the sample would skew toward stronger applicants and/or people who were admitted and therefore more likely to share their stats. On top of the bias issues, it's also easy to misclassify fields like program, university, and even term/status. As we saw in last week's module, entries are inconsistent because everything is typed in by the applicant, so program and university names vary a lot (spelling, abbreviations, formatting). That means my database and queries can misclassify records, and my results depend heavily on how well the data is cleaned and standardized.