Web Scraping Limitations: Grad Cafe Web Scraper

One major limitation of scraping unfiltered raw data from Grad Cafe (or any website for that matter) is that outlier data can be entered without any sort of repercussion. For example, while visually examining some data, some college names were missing, spelled wrong, or were simply not colleges at all. There also were not ethical "rules" involved in applicants submitting their data, meaning they could simply make things up. If the GRE quantitative reasoning score of the sample data set provided was 165 when the national average is 157, then there is clearly a discrepancy in the data being uploaded versus reality. This skew in the data is likely because of people applying to programs that require higher scores than average. Another reason for this discrepancy could simply be because people who are scoring lower/not getting into these programs may not even be uploading data in the first place to Grad Cafe.

When I carried out my own analysis, nothing stood out too much aside from my custom question about the number of applicants per degree type. What surprised me about this was the number of PhD candidates applying to programs: 19,842 records out of 30,000 were applying to PhD programs. I would have expected many more potential graduate students would be applying rather than PhD students. There were a fair amount of graduate applicants: 8,634/30,000, but for there to be more than double the amount of PhD candidates was surprising for me.