Module 3 — Limitations of Anonymous, Self-Submitted Data

The data from GradCafe is submitted anonymously by students. It is not verified so the input can be whatever the user enters. The people who post are not providing us a complete view and may have report garbage data, or data that is very strong, or very weak. They also can round off their scores or GPAs not consistently and make typos or rounding errors. The submissions are free text so names of programs and universities and even status may have different spelling of format and will likely be misclassified unless they are normalized by a tool. So aggregate statistics should be viewed as those of the people who posted and not reflecting all the real applicants.

Some analysis can look unusually high compared with benchmarks because of who opts and chooses to post and what they elect to report - e.g., a higher than expected average GRE Quant score can come from if high-scoring applicants are more likely to post decisions, if most posters come from programs/subfields where GREs are relatively high relative to other metrics, or if, say older entries from times when GRE was required. One additional factor is records that are not there. Sometimes applicants might have low scores and choose not to report them (so the average is higher because of who chose to provide resposnses on the site. Additionally, we should remember that the inconsistent formatting like numbers outside of the ranges for tests, means that this data is to be used as exploratory and for testing theory, and not to represent of provide 100% accuracy for all students.