Module 3 – Limitations of Anonymous, Self Submitted Data

When I started looking at the results of my SQL queries, the numbers did seem high to me. The average GRE Quant score in my dataset was 164. This is way higher than the national average (I usually saw a national average in the 153-158 band depending on the year so statistically significant!). So… I wasn't super shocked because GradCafe isn't random sample of all students. It is a biased, self selected and self reported sample. Presumably students that get super good scores and students that are applying to top 5 programs like MIT or Stanford are way more likely to be on the internet talking about their application results. Students that didn't get in or didn't get high scores usually stay quiet which makes the entire dataset drift towards the "successful" end of the spectrum. 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 reported 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.

Another limitation I noted was the inconsistency in detail that students can enter different things into their university field. In my database I had to do a comparison of the "original" university names against the "LLM cleaned" university names just to even get a proper count for CMU/Carnegie Mellon. And without that cleaning step that I did, my SQL queries would have missed dozens of applicants who typed CMU or made typos in their records. This is why while, anonymous crowdsourced datasets are interesting for getting rough trends (e.g. which programs are super popular), the data has typos, noise and is biased and we should not treat this data as official stats. It also highlights why cleaning data is as important as analysis. 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 responses 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.