Limitations

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!). Like… I wasn't super shocked because GradCafe isn't random sample of all students. It is a biased, self-selected 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.

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 noise and biased and we should not treat this data as official stats. It also highlights why cleaning data is as important as analysis.