Analyzing anonymously submitted data, such as the entries on GradCafe, presents several limitations that can affect the accuracy and representativeness of the results. First, the self-reported nature of the data introduces the possibility of selection bias. Students who received positive outcomes may be more likely to share their results publicly, whereas those with rejections or lower scores may abstain. Additionally, if someone were to be bored, they could include some nonsense entries that would further skew collected data. With the large number of entries being collected this can introduce multiple inconsistencies and overall incompleteness. This lack of standardization complicates the cleaning process and can skew averages or obscure trends that might otherwise be evident in a verified dataset. Personally throughout this assignment, I think the one datapoint that surprised me the most was how high the average GPA acceptance was. A 3.7 in my eyes is almost near perfect to me, but I think it further proves my point on selection bias.

One particularly surprising result was the observed average GRE Quantitative Reasoning score of nearly 165, compared to the national average of 157 in 2023. This discrepancy could stem from the overrepresentation of technically strong applicants—particularly those applying to engineering, computer science, or other quant-heavy programs—who are more likely to use platforms like GradCafe. These users might also be more inclined to report their outcomes, further skewing the sample. Ultimately, while such platforms can offer valuable insight into the admissions landscape, the data should be interpreted cautiously and not as a comprehensive or unbiased reflection of applicant pools. For more robust understanding, data should be scraped from multiple sources.