After completing the web scraping and database creation assignment, I noticed that an inherent limitation with leveraging a data source based on anonymously submitted graduate school application data is that we’re relying on those who are voluntarily reporting and/or aware of the website. We’re also forced to assume what is reported is true, since the website lacks a means of data validation and verification. Also, there are variations in regard to the scale of numerical values like GPA (some go by a 4.0 scale while others can be 5.0). From a data quality perspective, we did as much as we could with handling missing fields, inconsistent formatting, and varying levels of detail across submissions for correcting incomplete or unstandardized data with the LLM. The questions I used to query were to see how sparse the data is by choosing a pretty popular program at a well-known university and seeing there was only 1 value. The other question was to see if there was a skew in GPA throughout all terms and degrees for a single university's PhD acceptances.

The average GRE quantitative score of 164.9 is quite high compared to the average of 157 showing the selection bias of using this data source. Maybe applicants with stronger qualifications feel more inclined to submit their scores or maybe they could also be making up numbers leading to a positive outcome bias. Someone could also just be publishing their best score. Some ways to work around issues with crowd-sourced public data are implementing features like confidence intervals, bias warnings, and benchmarks against official statistics so that we’re highlighting any potentially skewed representations.