The inherent limitations of performing analytics on a platform like Grad Cafe stem primarily from its anonymous and unverified nature. Because the database relies on crowdsourced entries, it is highly susceptible to unreliable data. Anyone with access to the internet can post any metrics they wish without having to provide a transcript for verification. This introduces unverifiable bias where users might inflate their scores to feel more competitive within the community. Furthermore, the lack of a standardized entry format leads to data integrity issues. For example, during my analysis, I found that discrepancies between raw data and LLM-generated fields often arose because the model struggled with non-standard abbreviations or typos that a human might catch but a machine misses, leading to mismatched results for elite school applicants. These human-induced errors can often be limited with the introduction of set choices to select from or adding lag to the system (a second pop-up window to verify that the input information is in fact correct).

The analytical responses were surprising at first, particularly the 50% acceptance rate for Fall 2025, which felt low at the time until I considered the possibility that students with multiple competing offers may only report their successes, or they may be accepted to programs they ultimately choose not to attend. However, the most striking difference is how these numbers diverge from national standards. While the national average GRE Quantitative score is approximately 157, my data showed scores closer to 165. This is likely caused by a concept I learned to be Self-Selection Bias. Students with lower scores are far less likely to post their results publicly while high-achievers are more motivated to share their success. Consequently, the sample pool we are scraping isn't a representation of all applicants, but rather a subset of the most competitive ones, which naturally skews the data toward the high end of the scale. Being unable to connect my scrape.py with the analytics also limited my ability to produce more accurate results due to the limited data pool. With the introduction of 50,000 more data points, we'll be able to paint a better big picture of the questions imposed.