Written Assignment Requirements

Having carried out this assignment, please write two paragraphs about the inherent limitations of carrying out analytics over anonymously submitted data items. Did the analytic responses surprise you? How is this different from standards? For example, the average GRE quantitative reasoning score was 157 for 2023-2023 and was nearly 165 for grad school entries submitted (see sample output). Why do you think that is? What might cause this to occur? Please place your essay into a file called limitations.pdf

The inherent limitation of carrying out analysis over anonymously submitted data is that there are issues regarding selection bias or false reporting. This can be a result of people inputting false data, or that stronger applicants, applicants with unusual outcomes, or specific programs are more likely to post their information, therefore skewing the data. There is also the issue that because it is prone to false reporting or limited reporting, there is no need to report parts of the applicant profile that may be critical. For example, it is likely that an applicant got accepted but did not submit their low GRE score, or vice versa, causing a skew in the data. This means that even with cleaning and LLM implementations, the inherent limitation of the website being open to the public and having minimal submission constraints makes the data have some issues. The analytic responses did surprise me in that the acceptance rate to these programs is around 26 percent, which also supports the idea that there is an over-representation of the extreme, where in reality the acceptance rates for master's and PhD programs are not exactly as shown on the website.

I believe that these inflated values are not a good representation of the whole population standard, since they are caused by non-random sampling and reporting behaviors that are more likely to be on the extreme. The average GRE quantitative reasoning score of grad entries being so high can be a result of selective reporting and bias toward reporting higher scores, since it makes the applicant look more competitive. It can also be because an applicant with a high score is more likely to input their scores to the website compared to someone with a lower score. This also explains why the sample output can show a GRE quantitative average near 165 while broader standards are closer to 157. Overall, I find the analytics are useful for spotting trends in posted results, but they should not be treated as exact population-level statistics as there are various sampling errors and biases that are inherent to the type of website GradCafe is.