Name: Arjun Bamba

Question: Having carried out this assignment, please write two paragraphs about the inherent limitations of carrying out analysis over anonymously submitted data items. Did the analytic responses surprise you? How does 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?

Conducting analysis on data submitted anonymously comes with many challenges including but not limited to: self-reporting bias, incomplete data, sampling/selection bias or outcome bias, non-representative sample, lack of standardization, duplicate/spam entries, and no verification. All these challenges impose limitations on the overall usability and reliability of the data. The data is prone to self-reporting bias since anyone can self-report their results and stats without any verification whether it's true. Because of that, we are also prone to incomplete data due to errors like typos or misreporting intentionally (ex: inflating stats). The data could also be incomplete because of key information being left out or vaguely labeled/referenced or entered into the wrong field - all of which makes it hard to effectively clean and analyze the data. Another indirect challenge is sampling/selection bias or outcome bias. Since our sample only contains data from the people that shared it on Grad Cafe, we need to be mindful that it may not be a good representation of the entire pool (due to reasons like only successful applicants sharing their results which can skew our analysis).

Our analysis did involve working through many of these challenges. As per the given example, the average GRE quantitative reasoning score was 157 for 2023-2023 and was nearly 165 for submitted grad school entries. This indicates a selection bias since it's evident that applicants who had stronger stats were more inclined to post their results than applicants who had slightly lower stats. In addition, the sample may also be non-representative since students from certain colleges/degrees may be more inclined or vocal in terms of sharing their results while students from other disciplines may not be as vocal. This is different for more standardized datasets that may be collected by universities themselves because standardized datasets (by universities) would have official verified stats where we know that the reported stats and outcomes are actually true and representative of all the applicants. Because of these many challenges, anonymous submissions should be thought of more as a brief, incomplete insight providing only a glance of the broader official data.