Jacob Wong

06/08/2025

Module 3 Writing Assignment

Prompt: *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 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?*

One inherent limitation of carrying out analytics over anonymously submitted data is the risk of introducing sampling bias. Sampling bias is a skew in data that occurs when a sample is not representative of the overall population. This essay prompt stated that "the average GRE quantitative reasoning score was 157… and was nearly 165 for grad school entries submitted." In this example, we see a positive skew of the data from Grad Cafe website because high-performing students are more likely to report data to the Grad Cafe vs low-performing students. Scraping data from a website like the Grad Cafe provides data analysts with little control over the sample, and thus, sampling bias is inevitable.

Another limitation of anonymous data is the incompleteness of individual submissions. In my analysis, I noticed that very few entries included their GRE scores. These scores can significantly change the results of a student's application. For instance, suppose that an accepted student has a below-average GPA but above-average GRE scores. If the student decides to submit their GPA but not their GRE scores to Grad Cafe, that student's entry will lower the average accepted GPA to that program without contributing the GRE average. In proper data analytics, ensuring uniformity across all entries is essential to achieve a holistic understanding of the dataset.