Written Assignment: Limitations of Anonymous Data Analysis

Author: Wei Liu
Course: EN.605.256.8VL.FA25 - Software Concepts
Assignment: Module 3 - Analysis of Grad Café Data Limitations
Date: September 13, 2025

Analyzing anonymously submitted data from platforms like GradCafe comes with several limitations that restrict the reliability and generalizability of conclusions. The most significant issue is **selection bias**, since the dataset is drawn only from individuals willing to voluntarily share their admissions outcomes. This produces a skewed sample that may overrepresent either highly successful applicants eager to showcase achievements or disappointed candidates seeking support after rejections. **Self-reporting accuracy** is another concern, as anonymous entries cannot be verified; contributors may misremember, exaggerate, or even inflate details such as GPA or GRE scores. In addition, **temporal clustering** of submissions—shaped by application cycles, admissions timelines, and shifting platform use—creates uneven representation across institutions and years. Finally, **demographic imbalance** is likely, since users of such platforms may not reflect the diversity of the overall applicant pool and may lean toward certain regions, socioeconomic groups, or academic fields more active in online forums.

The results from this dataset also produced outcomes that differ noticeably from official benchmarks, particularly in standardized test performance. The average GRE Quantitative score of **161.87** was substantially higher than the 2022–2023 national average of **157**, a difference large enough to signal systematic bias rather than chance. Several factors likely contribute: **self-selection bias**, with stronger applicants disproportionately represented; **score inflation**, whether intentional or from optimistic recollection; and **platform demographics**, which may draw candidates from competitive academic environments or regions with intensive test preparation. The unusually high proportion of **international students (53.53%)** in the dataset further suggests a non-representative sample with different preparation patterns. In addition, **social proof effects** may encourage contributors to align their reported scores with the high-performing community they observe, reinforcing inflated results. Together, these dynamics highlight the importance of understanding the origins of crowdsourced, anonymous datasets and applying statistical adjustments before making inferences—especially in contexts like graduate admissions, where accurate comparisons to official standards are crucial for guiding applicants.