Limitations – GradCafe Analysis (Module 3)

This project relies on admissions data originally scraped from the GradCafe website. Although the final Module 3 database and analytics pipeline function correctly, the overall analysis is subject to several important limitations.

First, the GradCafe platform is based entirely on anonymous, self-reported information. Users voluntarily post their application results without verification, which introduces substantial self-selection bias. Applicants with extreme outcomes—such as acceptances, rejections, or unusually high test scores—are more likely to post than those with average results. Consequently, the dataset cannot be considered a statistically representative sample of the overall graduate applicant population.

Second, the data is highly unstructured. Most posts consist of free-form text rather than standardized fields. Critical information such as GPA, GRE scores, nationality, decision status, and application term must be inferred using regular expressions and heuristic parsing. These extraction methods are imperfect and frequently fail when users write in unexpected formats, use abbreviations, or omit details entirely. As a result, many records contain missing or incorrectly interpreted attributes, and averages are computed only over the subset of rows where values could be detected.

Third, this submission depends on an instructor-provided cleaned dataset rather than on the student's original Module 2 scraping output. The original Module 2 dataset contained substantial structural problems, including incomplete fields, inconsistent formatting, and missing records, which made reliable analysis impossible. Following instructor guidance, the cleaned dataset provided by Liv d'Aliberti was used as the single source of truth for Module 3. Therefore, the accuracy and completeness of all Module 3 results depend entirely on the quality of that external dataset rather than on the original scraping pipeline.

Fourth, there is significant temporal bias in the data collection process. The scraper used in Module 2 collected records using a "newest first" approach, meaning that the dataset is heavily skewed toward the most recent admissions cycle (primarily Fall 2026). Older years are underrepresented, making longitudinal comparisons unreliable. Acceptance rates and distributions by term therefore reflect recent posting behavior rather than true historical trends.

Fifth, program and university names are highly inconsistent across posts. Applicants frequently use abbreviations or informal language such as "JHU," "MIT," or "CS," requiring extensive normalization. Even with local language-model processing, some program classifications remain ambiguous or incorrect. This affects analyses that rely on matching specific universities or degree types.

Finally, many potentially relevant dimensions—such as funding information, multiple degrees, conditional admissions, or application context—are not captured in structured

form and therefore cannot be analyzed. The resulting database necessarily simplifies complex real-world outcomes into limited categorical fields.

In summary, while the Module 3 system correctly demonstrates database loading, querying, and web presentation, the underlying data should be interpreted only as an exploratory sample of GradCafe posts. All numerical results are approximate and descriptive rather than definitive measures of real-world graduate admissions patterns.