Working with anonymously submitted GradCafe data showed several limitations that affect how reliable the analysis can be. Since the data is self-reported, there is no way to verify whether entries are accurate, complete, or consistent. Some users may only report certain results, which can introduce bias into averages and percentages. The dataset also contains many inconsistencies in how people describe universities, programs, and degrees. For example, computer science programs may appear as “CS,” “CompSci,” or “Computer Science,” and universities may be written in multiple ways. Because of this lack of standardization, extra preprocessing was necessary before meaningful queries could be run. Missing values and inconsistent formatting also mean that results should be interpreted as general trends rather than exact statistics.

Some of the analytic results differed from expected benchmarks, such as higher average scores compared to official statistics. One possible reason is reporting bias, where stronger applicants or accepted students may be more likely to share their outcomes online. Another issue is that the sample is not random, so it may not represent the entire applicant population. For Question 7, I chose to use normalization instead of regex to handle inconsistent naming because it was simpler and easier to maintain. Normalization allowed me to standardize text by converting it to lowercase and matching known keywords, which reduced complexity while still handling common variations like “CS” or “Computer Science.” This approach felt more readable and practical given the scope of the assignment.