Limitations of collecting data from grad cafe

Challenge 1: DATA COMPLETENESS AND MISSING VALUES

One of the primary challenges I faced during this assignment was the substantial number of missing values in the dataset. Out of 32,000 scraped entries, a large portion lacked complete information in several fields. For example, approximately 53% of entries (16,949 out of 32,000) did not include comments, and GRE score fields were particularly incomplete. When I first scraped the Grad Cafe website, I observed that GRE Quantitative and GRE Verbal scores were largely missing from the main listing pages. The site only displays basic information in the results list, while more detailed metrics are available on individual entry pages. To obtain GRE data, I adjusted my scraper to access each individual result page, which took over 40 minutes for just 4,000 entries due to rate limiting. Despite these efforts, only about 2.5% of entries included GRE Quantitative scores (105 out of 4,000), and about 6% included GRE Verbal scores (245 out of 4,000). This limited data set complicates statistical analysis, as averages calculated from such small samples may not accurately reflect the broader applicant population. The high proportion of missing values likely results from both the optional nature of self-reported fields and the recent trend among graduate programs, especially after COVID-19, to make GRE scores optional or eliminate the requirement.

Challenge 2: SELECTION BIAS AND INTERPRETATION CHALLENGES

Beyond data completeness, I found that analyzing anonymously submitted data from Grad Cafe introduces significant methodological challenges, especially about selection bias. Individuals who submit their application results are not a random sample; they tend to be more engaged with the application process and often have stronger profiles or notable outcomes. This is reflected in my findings: the average GRE quantitative score in the dataset was 166, compared to the national average of approximately 157 reported by ETS, a difference of 9 points. Similarly, the average GRE Verbal score was 160, which is considerably higher than the national average of 151. These results suggest that high-achieving applicants are more likely to report their scores, while those with lower scores may be less willing to share them. The average GPA of 3.76 in the dataset also exceeds that of typical applicant pools. Furthermore, the data does not capture applicants who never

submitted results, withdrew applications, or chose not to disclose negative outcomes. These biases imply that prospective applicants relying on Grad Cafe data to assess their competitiveness may develop unrealistic expectations. Therefore, the data should be viewed as descriptive of a self-selected, high-achieving subset rather than representative of all graduate school applicants.

Reference:

These answers were originally written by me, and I used AI to polish the paragraph just to make sure the statements sound precise and professional.