

Limitations of Analytics of Anonymously Submitted Data Items
Modern Software Concepts in Python
Emily Hammer (ehammer5)

One of the limitations for this dataset is that a single person would often submit multiple entries, one for each school that they had applied to and/or heard back from. While the skew would be small, there is still a bit of distortion on the aggregate measures (e.g. average GPA) because a single person can submit the same score multiple times across several entries. If this dataset was non-anonymized, it would be easier to have two separate tables: one of applicants and one of applications. This would allow us to calculate measures over the aggregate applicant population (e.g. average GPA where one GPA is submitted per applicant) and data for all applications (e.g. total number of applications where one applicant may have multiple applications).

Another limitation is the inherent bias of self-reporting. First of all, only certain types of applicants are likely to submit their data, and resulting analytics will reflect that bias instead of true numbers reported by universities. Additionally, for those applicants who do submit their data, they may alter it to make themselves appear better or worse depending on what the situation calls for. For example, they may not submit results in which they were 'Rejected,' only 'Accepted' or 'Waitlisted.' Or they may include in their 'Comments' exaggerated or untrue details that, if we were doing a sentiment analysis, could distort the outcomes of such results. While self-reporting can result in a large number of data points, it's important to keep in mind that it is not as reliable as primary-source reports, like an admissions report from a university. This unreliability is what can lead to an official average being 157, but a self-reported average being 165.