Limitations

As I worked on the assignment, I have observed several possible limitations when it comes to carrying out analytics over anonymously submitted data items. The first limitation is the uncertainty of data input accuracy from the webpages that we scrape the data. Since users are anonymous, we would not be able to know if one user inputs more than one record, that would definitely affect our analysis as the data is skewed due to unreal data input. The second limitation is the possibility that users have typos in their inputs. At first, when I ran my program, for both question 4 and 6, I received the average GPA to be more than 4.00, which surprised me and lead me to inspecting the data in the data base. I found several occasions when GPA were input as 389 instead of 3.89 or 378 instead of 3.78. These typos inflated the average GPA calculations and provided less accurate answers. Thus, in my program, I exclude those from being included in the calculations. The third limitation is the availability of data to constitute a good sample. For example, not all students who applied to universities report the status to GradCafe.com. Also, there is also the chance that certain schools receive more reports than others. These can lead to a distort in the randomness of the sample data.

The limitations mentioned above are from the external source, we also have to be aware of the internal limitations. As we scrape the data and parse it into standardized format, it is important that we are able to parse the raw data into well fitted format for further analysis. If we are unable to ensure such standardization in our parsed data, it may lead to errors that skews our analysis. For example, in my data set, I found that Georgetown University is reported as “Georgetown University” in the raw data, but in the llm_generated_university, it is reported as “George Town University”, and this difference caused the answers to my question 8 and 9 to be different at first. I changed the search term in question 9 to “George Town University” and received an answer that is similar to that in question 8.