Using anonymized and user-generated data has many limitations. One of the first limitations is the cleanliness of the data. We spent a lot of module 2 just normalizing both the program and university names, but they still are not 100% clean. There also is no way to validate the data or to ensure that everything is entered. In some cases, there are GPAs and it wasn't entered in other cases. The same issue with GRE scores as well. There's also cases where duplicate entries might have been entered, but since it's all anonymized, there is no way to 100% validate it. Additionally, since the data is voluntarily entered, there may be a bias in terms of the individuals who are likely to be entering the information.

In terms of surprising results, it likely has to due with who is entering the information. In the case where the average GRE quantitative reason score is 165 versus the average of 157, it could be due to the fact that users with higher scores might be more likely to have entered their information versus users with lower scores. There's also no way for us to validate that the scores that users entered are indeed the scores they received. As a result, the self-reported data might not be fully representative of the full population so we need to be cautious when trying to draw any conclusions.