Inherent limitations of anonymous datasets

A dataset consisting of public and anonymously submitted data has a lot of limitations in its trustworthiness. While the data may be more readily available, it suffers from a lack of controls over the allowed inputs. Therefore, a significant chunk of the dataset is potentially wrong, malformed, or even maliciously incorrect. Even worse, it's extremely difficult, if not virtually impossible, to know for certain *how much* of the dataset is correct or to what degree it is corrupted.

Perhaps the biggest problem of anonymously submitted data is a complete lack of proper sampling. The users that submit data are likely incentivized to do so for one reason or another, and this leads to a catastrophic sampling bias. For example, a rejected student might submit their data with an inflated test score as a means to defame the system somehow. Additionally, the submitters who are aware of- and who take the time to seek out and use- the online system are likely entrenched and do not reflect the majority.

Thus, while an interesting exercise, it's extremely difficult to draw any reliable conclusions from the dataset scraped from anonymous online sources.