Dylan Breen
Modern Software
Module 3
9/16/2025

The Psychology of Data Gatherers

The inherent limitations of using a self-submit type data source can be broken down into several principles. One is the weakest link/least common denominator principle. This means that your data will only be as good as your worst user. In effect, you will have to create conditioning/validation/cleanup procedures designed to capture and either fix or eliminate data that is "dirty" based on the worst user (not a personal value judgement, just a descriptor). In essence, best practice involves creating a system/application for the average or, when appropriate, worst user. This lowers the risk of and accounts for junk data and bad user habits, which in turn increases data quality and integrity.

Much like a self-reporting survey in such areas of psychology, the other limitations include the accuracy of reported data. Without rigor of practice and the inability to independently verify the source, self-reporting jeopardizes the veracity of the data itself. The example in our assignment was given of the actual averages of gre (which can be independently verified through testing data made available by ETS) was much lower than the self-reported values on the grad café site. There are multiple reasons this could be the case. One could be the users themselves are inflating their datapoints for various potential psychological reasons. Another could be that the people who actually self-report already select a subset of the actual population data which biases the data itself via the selection medium. There are multiple other unknown reasons, which shows you would also have to run additional analysis to solve that question when using a self-report, anonymous site that can’t be independently verified or doesn’t have rigorous controls and practices to ensure the data quality. Self-submission if not tightly controlled with data validation and governance that sanitizes the data at entry means even with the best cleaning mechanisms, there is risk that data will be missed and take extensive work to fix or eliminate from a dataset. Additionally, data you exclude may affect the integrity of the analysis later depending on the type of data and volume. Entry point is often a missed opportunity; however, many software designs have this in mind and allow extensive governance of user input. Staying away from strings when possible does help.