Data Mining / Institutional Data Checklist

Instructions

Use this checklist as you develop your project. Although it is a checklist, the purpose of its use is not to “check” every item. Rather, the goal of this tool is encourage thoughtful reflection and consideration of critical aspects of the implementation of a data mining / institutional data use project. Checking all the items on this list will not guarantee a perfect project, but the discussions you have regarding each item will help you improve your project. Finally, the tool is not designed to be used alone - it is best discussed with a team.

Clear Purpose

- □ Data are available to address the topics or issues of interest.
- □ What we want to learn from the project is clear and specific.
- □ We have the tools and expertise to perform these needed analyses or can enlist the help of someone who does.

Identifying Data

- □ Search for usable, meaningful data to address our project’s purpose includes both internal and external sources.
- □ Data quality (accuracy, relevance, trustworthiness) is carefully considered when selecting sources of data.

Respecting Rights of Participants

- □ If sensitive data are used, IRB approval is received.
- □ If IRB approval is not needed, adequate protections are still used to ensure the privacy and rights of those individuals who are included in the data set (such as security of the data file, confidentiality statements from those who are working with the data, limiting the data file to those variables of direct relevance for the project, and removing identifiers as early as plausible).
- □ Care is taken to avoid identifying individuals through extreme disaggregation of demographic variables (e.g., student from XXXX, of gender XXXX and race XXXX, majoring in XXXX, and participating in Division 1 sport XXXX).
- □ Don’t be creepy. Use caution when using results to contact students or take other actions with identifiable data used in data mining.

Data Cleaning

- □ Before analyzing any data, data are checked for accuracy, missing data, and typos.

- □ A clear data codebook is developed.
- □ Clarifications on the meaning of variables and codes are received if needed.
- □ Data are formatted in a manner that is suitable for your selected analysis software (e.g., Excel, SPSS, Tableau, etc.).
- □ Descriptive analyses are run and unusual and unexpected results are investigated.

Running Analyses

- □ If running multiple analyses, be cautious of the group-wise error rate (i.e., the more statistical tests you run the more likely you are to find a statistically significant difference by chance).
- □ Keep track of the analyses you've run and keep a log book or notes about what you are finding.
- □ Select the best tool for the job.
- □ Running a sophisticated statistical analysis does not guarantee that you will find something interesting. Use the best statistical approach for the project.
- □ Ask for help if you need it.

Reporting Results and Next Steps

- □ "Less is more." You don't need to report results from every analysis that was run. Rather, focus on those that best help answer your question(s).
- □ Know your audience and develop a report that is understandable by that audience.
- □ Consider the use of graphs and other data visualizations (such as infographics).
- □ Recognize the limitations of data mining and analyzing institutional data. Connect your findings with other findings (such as the results from surveys or focus groups).
- □ Identify next steps to make use of your findings.