




Five Key Questions about Public Use Files

Posted by [China Layne](#)

Find me on: 

Read more about me: [Biography](#)

WHAT IS A PUBLIC USE FILE?

A public use file (PUF) is a dataset, generated from survey, administrative, or mixed data collection methods, that is suitable for use by public researchers. PUFs allow data collectors to provide the public with access to high-quality, easy-to-use data while maintaining respondent privacy.



WHO NEEDS ONE?

Any organization that collects data for research and analysis should consider creating and providing access to PUFs. PUFs bolster research transparency and can be created by an experienced organization with minimal effort.

WHAT ARE THE BENEFITS OF OFFERING ACCESS TO PUBLIC USE FILES?

Offering access to public use files provides several benefits for research organizations:

1. Enables an organization's research findings to be replicated and verified.
2. Increases the usefulness of an organization's data collection efforts by supporting future, secondary analysis of the data.
3. Establishes the organization as a "go-to" destination for resources for specialized research.

WHAT ARE THE KEY PARTS OF A HIGH-QUALITY PUBLIC USE FILE?

A high-quality public use file includes three key parts:

1. A high-quality dataset, free of data errors,
2. A dataset that limits disclosure of respondents' sensitive information, and
3. Adequate documentation for using the public use file.

HOW DOES SUMMIT CREATE PUBLIC USE FILES?

Summit's approach to creating high-quality public use files starts with reviewing the analysis dataset to ensure that it is **free of data errors**, such as mislabeled data, invalid values, and internally inconsistent variables.

Next, Summit thoroughly examines the dataset to ensure that it **limits disclosure** of respondents' personally identifiable (PII) and sensitive information. Summit employs statistically sound methods to limit disclosure of respondent information, such as aggregating values of variables with very few respondents, limiting availability of geographic and sampling information in the dataset, and aggregating values of sensitive variables (e.g. earnings and wages).

Finally, Summit creates documentation that provides **information for correctly using the public use file**, including: (1) a complete codebook, (2) tabulations of weighted estimates of the key variables with measures of error, and (3) a research note outlining proper weighting and variance estimation procedures for the data and information on any variables specifically constructed for the analysis. Summit has used this approach for creating and evaluating public use files for the Department of Labor's (DOL) Chief Evaluation Office (CEO) and has developed a standardized checklist to guide this work.

ADDITIONAL RESOURCES

[Summit's Program Evaluation Team](#)

[Confidentiality and Data Access Issues Among Federal Agencies](#) (published by Federal Committee on Statistical Methodology)

[Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology](#) (published by Federal Committee on Statistical Methodology)

***Meet the Program
Evaluation Team >>***

Topics: [Data Analytics](#)