Tutorial 10 - Datasheet*

Datasheet for US General Social Survey variable/dataset premarsx

Russell Luchin

March 20, 2024

Table of contents

*Repository available at: https://github.com/rluchin/tutorial10

1 Motivation

1.1 Purpose

The dataset was created to track social opinion on the notion of premarital sex in order to gauge how "moral", or socially acceptable, the act is. The variable is not intended to fill a specific gap, but to assist in creating a comprehensive picture of what Americans consider morally sound in relation to sexual intercourse.

1.2 Creator/Researcher

Data is collected by the National Opinion Research Center, a part of the University of Chicago but is funded by the National Science Foundation on behalf of the United States' Federal Government.

1.3 Funding

Collection was funded by the US government through grants directly issued to NORC at the University of Chicago.

1.4 Additional Comments

Research is conducted by the University of Chicago but all the survey, and by extension this variable, is collected at the behest of the US government.

2 Composition

2.1 Dataset Representation

Each row of the dataset is an individual collected variable from the GSS. Combining these variables together creates a comprehensive dataset where each column is a variable and row a response to the tracked variable of the column.

2.2 Amount of Instances

Approximately 72,000 instances.

2.3 All possible instances?

Yes, all historic responses to this variable from the GSS are recorded in the dataset, for each instance it occurred in the survey.

2.4 Instance Composition

Each instance contains the response of the 4 possible types: “Always wrong”, “Almost always wrong”, “Sometimes wrong”, “Never wrong”

There is no label associated with each instance unless the data is constructed with additional GSS variables.

All identifying information is missing from the dataset except for age and year of survey.

Instances are provided an ID which identifies an unspecified relationship between respondents/surveys.

There are no data splits.

The responses “Inapplicable”, “No answer”, or “Skipped on web” create noise when tracking responses to the premarsex variable.

Data is self contained.

Age, region, and sex of the respondent can be tied to the premarsx variable, but no explicitly identifiable information can be viewed.

Dataset contains responses to a subject that can be considered controversial in some communities, but is largely uncontroversial as a whole due to the nature of it being collected as part of a social opinion survey.

Dataset does not contain sub-populations unless multiple GSS variables are pulled to construct a new dataset.

It is impossible to identify individuals based on the variables in this dataset.

No sensitive data is contained within this dataset.

3 Collection Process

3.1 Acquisition

Data was acquired primary from in person interviews from a random sample. Some years of data collection were collected through online surveys. Respondents were chosen based on random sampling conducted by NORC.

Human curation and online survey tools were used to collect the data.

3.2 Sampling

Data is not a sample.

3.3 Parties Involved in Collection

The interviewee, the interviewers and the stakeholders of the institutions conducting the survey (NORC, National Science Foundation, and the US government) were involved in the collection of the dataset.

3.4 Data Collection

Data was collected over 50 years (1972 to 2022). Dataset was updated every time the variable was collected as part of the GSS. In general, this would be every 2 years.

Dataset was acquired directly from NORC, using their GSS explorer tool.

Individuals were notified and needed to provide consent to be interviewed. Individuals were notified and agreed to have their data collected as part of the survey, in most cases they were notified in person by the data collection specialist/interviewer themselves.

Individuals consented to the collection of data as part of the survey. Due to the nature of in-person interviews, no evidence can be provided other than speculation as to how the GSS interviewers obtained in-person consent from the interviewee.

4 Processing/Labelling

4.1 Cleaning

Data was not cleaned.

Dataset consists of only raw data. Dataset is attached at GitHub repo linked at beginning of datasheet.

5 Uses

Dataset has been used by NORC to graph trends and key insights from data. Key insights can be found here.

There is no repository that links to all uses of this dataset.

Dataset can be used to graph and analyze trends that involve the tracking of the premarsx variable. This includes cross-referencing with age, region, or gender data.

Nothing about dataset collection/processing affect future uses.

Dataset can be used for any task related to premarsex variable.

6 Distribution

Data is currently available only through GSS data explorer, and will continue to be distributed this way barring a major procedure change.

Dataset is copyright protected and may not be disseminated anywhere except from the GSS data explorer tool.

7 Maintenance

Dataset will be maintained by NORC at the University of Chicago.

GSS website curator can be contacted at commhelp@norc.org.

Dataset is updated automatically by NORC at the University of Chicago roughly every two years.

There are no limits on the retention of data. As the data comes from regular surveys, the data from each year will be available continuously as part of the complete dataset.

There is no mechanism to extend or augment the dataset without being a part of NORC due to the nature of how the data is collected.