READ ME: GSS Teaching SURF and supporting documents
Kate Prickett || kate.prickett@vuw.ac.nz

The following READ ME file provides information on the 2018-2021 GSS Teaching SURF dataset and supporting teaching documents.

2018-2021 GSS Teaching SURF dataset
The 2018-2021 General Social Survey (GSS) Teaching Synthetic Unit Record File (SURF) is a dataset consisting of 1,997 synthetic respondents that, when variables are aggregated, produce population-level estimates for people aged 15 years and older in New Zealand in 2018 and 2021 similar to what would be estimated if researchers were using the GSS microdata.

The dataset request and coordination was led by Kate Prickett and Phillip Worthington at Victoria University of Wellington. Resourcing support for the 2018 data came from the libraries of Victoria University of Wellington, University of Auckland, Auckland University of Technology, University of Canterbury, and University of Otago. Resourcing support for the 2021 data was provided by the Victoria University of Wellington library.

The dataset was developed by Statistics New Zealand (StatsNZ).

The dataset consists of 1,997 unit records (997 in 2018, 1,000 in 2021) and 40 variables. Variables consist of a range of sociodemographic factors (e.g., age, sex, ethnicity, educational attainment, family structure, labour force status, household income, disability, home ownership, migrant status) and wellbeing indicators (e.g., life satisfaction, self-rated physical health, mental wellbeing, housing conditions, residential mobility, feelings of loneliness), as well as other behaviours and experiences (e.g., trust in government, voting behaviour, experience of discrimination, access to public transit and green space, being a victim of crime). A survey frequency weight is included.

Threedatasets are provided:
- GSS_SURF_2018_2021.csv/.dta: This dataset is best as an entry-level dataset. Some variables have been recoded for simplicity, and missing data, "don't know" and "refused to say" responses have been imputed with values.
- GSS_SURF_2018_2021_with_missings.csv/.dta: This dataset is best as an entry-level dataset. Some variables have been recoded for simplicity, and missing data, "don't know" and "refused to say" responses have been imputed with values.
- GSS_SURF_2018_2021_ORIGINAL.txt: This is the original dataset provided by StatsNZ. It contains many non-numeric values that need to be converted to numeric values, "don't know" values, etc. This is a good dataset for more advanced students that could benefit from working with how survey data typically look, including a set of replicate survey weights.

Dataset codebooks
One codebooks is provided:
- GSS_SURF_2018_2021– Codebook.xls: Codebook that can be used with the entry-level dataset and dataset with missingness retained.

Supporting documents

Dataset construction documents
- GSS SURF – simple dataset creation.do: This is a Stata .do file that calls in the original StatsNZ SURF, conducts recording (e.g., changing character values to numeric) and adds value labelling. Some variables are collapsed for ease of use (e.g., some top-coding, collapsing of categories), and "don't know" and "refused to say" answer options imputed with values. These files contain annotation for understanding recode decisions. The code creates the final working dataset "GSS_SURF_2018_2021".
- GSS SURF – simple dataset creation with missings.do: This Stata .do file creates the "with missings" dataset, retain "don't know" and "refused to say" responses (i.e., does not impute these values).

Teaching and assessment materials
These teaching materials are aimed at tertiary students who haven't engaged with statistics in many years or have a low statistical baseline to begin with. They will also have limited experience using statistical formulas in Excel or creating charts from derived statistics. The focus is on application and policy translation of findings.

The teaching materials are designed to work with the 2018 data only, but can be adapted to include both the 2018 and 2021 survey waves.

Creating and presenting bivariate statistics
- Workshop – Bivariate statistics.docx: Workshop instructions for creating bivariate statistics and presenting them in charts. Conducting Chi2 and t-tests.
- Workshop data – Bivariate statistics.xlsx: Data for the bivariate statistics workshop.
- Bivariate statistics assignment.docx: Policy memo writing assignment that assesses skills developed in the bivariate statistics workshop.

Conducting ANOVA/OLS linear regression
- Workshop - ANOVA simple linear regression in Excel.docx: Workshop instructions for running a simple linear regression (one independent variable) and creating dummy variables from categorical variables and including them in the model.
- Workshop data - ANOVA simple linear regression.xlsx: Data for the simple linear regression workshop.

Conducting and presenting OLS multivariable linear regression
- Workshop - Multivariable regression in Excel.docx: Workshop instructions for running a multivariable regression model and including interaction terms in the model.
- Workshop data - Multivariable regression.xlsx: Data for the multivariable regression workshop, along with answers to workshop questions.
- ANOVA-Simple linear regression and multivariable regression assignment.docx: Policy brief writing assignment that assesses skills developed in the ANOVA/OLS linear regression and OLS multivariable linear regression workshops.