UK Data Service

Tools for checking numeric data: QAMyData and sdcMicro

Cristina Magder Data Collections Development Manager
Anca Vlad Research Data Services Officer
Hina Zahid Senior Research Data Officer

SSHOC Open Science and Research Data Management Train-the-Trainer Bootcamp

10 May 2021

Photo by carlevarino on Unsplash

Cleaning data

Getting to know your data

- check structure and documentation to find issues
  - incorrect, missing, inconsistent values
  - check for unanticipated/accidental disclosure risk

Tools can:

- flag issues to enable a machine or human to resolve the problems
- be deployed
  - as a service for self deposit repository, eg DataVerse for a submission health check
  - into data publishing pipelines

QAMyData Tool

- UK Data Service developed a light weight, open-source tool for quality assessment of research data

- A ‘data health check’ tool that identifies the most common problems in data submitted in disciplines that utilise quantitative methods

Supports:

- researchers; creators and users to better appreciate how to respect/achieve/use high quality data
- data reviewers and publishers

QAMyData Checks

File checks

- File opens
- File name check

Metadata checks

- Missing variable labels
- Invalid variable names
- Missing value labels for defined missing

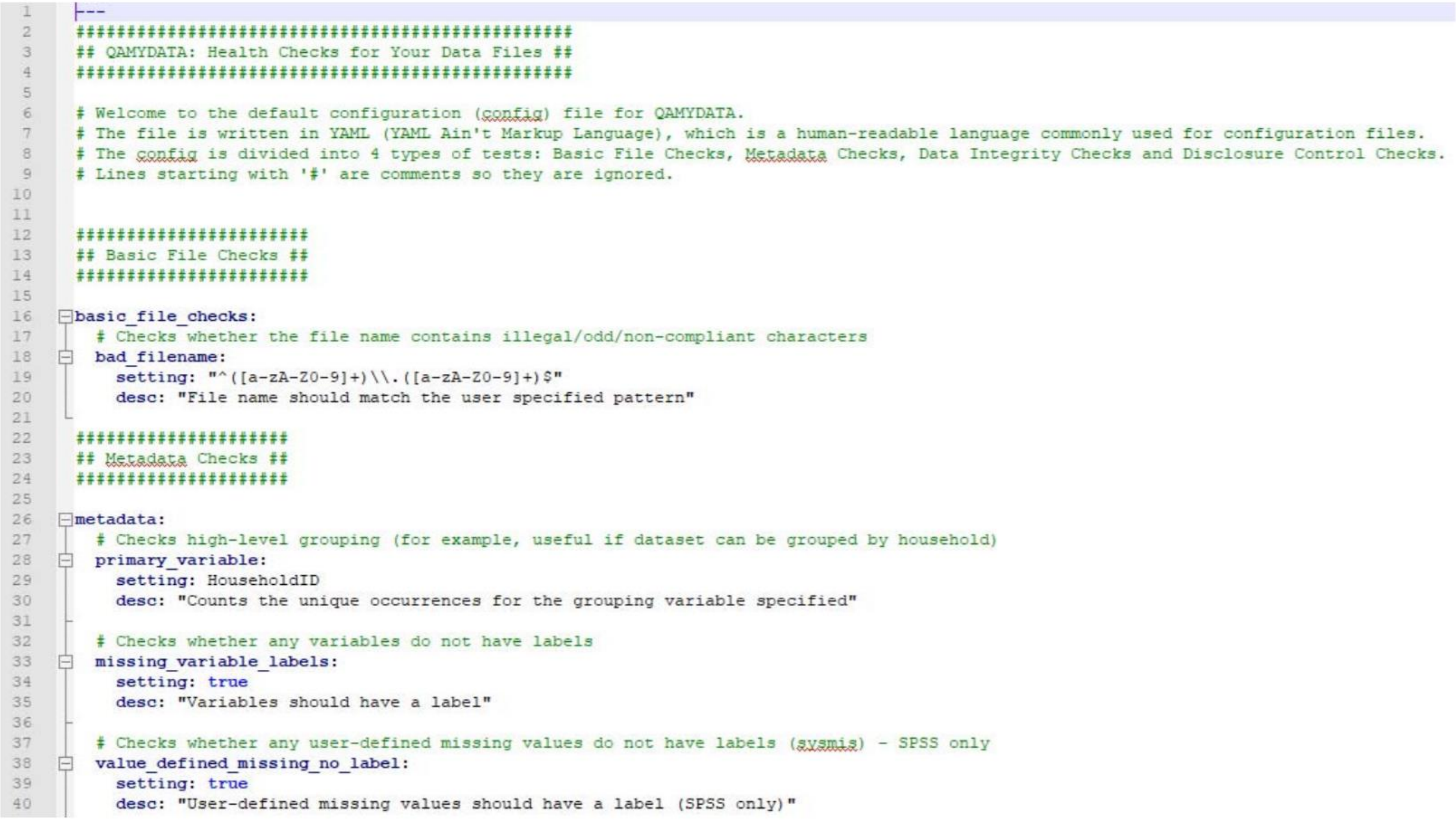
Data Integrity Checks

- Number of numeric and of string variables
- Odd characters
- Spelling mistakes and truncation
- Empty variables/missing information(system/defined missing)

Disclosure Checks

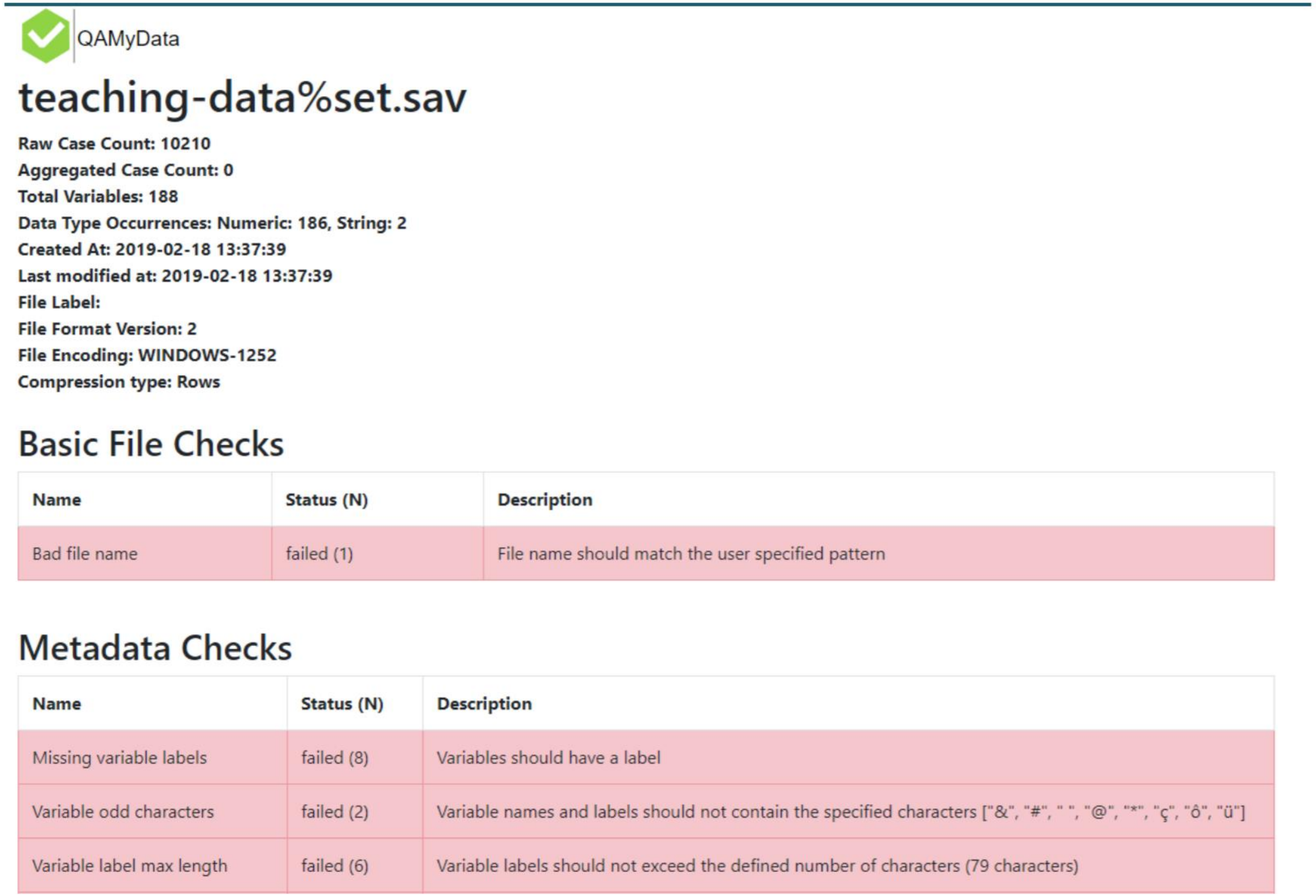
- Direct identifiers
- Unique values in continuous and categorical variables

Configuration for Tests



```yaml
---
#################################################
## QAMYDATA: Health Checks for Your Data Files ##
#################################################

# Welcome to the default configuration (config) file for QAMYDATA.
# The file is written in YAML (YAML Ain't Markup Language), which is a human-readable language commonly used for configuration files.
# The config is divided into 4 types of tests: Basic File Checks, Metadata Checks, Data Integrity Checks and Disclosure Control Checks.
# Lines starting with '#' are comments so they are ignored.


#######################
## Basic File Checks ##
#######################

basic_file_checks:
  # Checks whether the file name contains illegal/odd/non-compliant characters
  bad_filename:
    setting: "^([a-zA-Z0-9]+)\\.([a-zA-Z0-9]+)$"
    desc: "File name should match the user specified pattern"

#####################
## Metadata Checks ##
#####################

metadata:
  # Checks high-level grouping (for example, useful if dataset can be grouped by household)
  primary_variable:
    setting: HouseholdID
    desc: "Counts the unique occurrences for the grouping variable specified"

  # Checks whether any variables do not have labels
  missing_variable_labels:
    setting: true
    desc: "Variables should have a label"

  # Checks whether any user-defined missing values do not have labels (sysmis) - SPSS only
  value_defined_missing_no_label:
    setting: true
    desc: "User-defined missing values should have a label (SPSS only)"
```

Output file



QAMyData

teaching-data%set.sav

Raw Case Count: 10210
Aggregated Case Count: 0
Total Variables: 188
Data Type Occurrences: Numeric: 186, String: 2
Created At: 2019-02-18 13:37:39
Last modified at: 2019-02-18 13:37:39
File Label:
File Format Version: 2
File Encoding: WINDOWS-1252
Compression type: Rows

Basic File Checks

Name	Status (N)	Description
Bad file name	failed (1)	File name should match the user specified pattern

Metadata Checks

Name	Status (N)	Description
Missing variable labels	failed (8)	Variables should have a label
Variable odd characters	failed (2)	Variable names and labels should not contain the specified characters ["&", "#", " ", "@", "*", "ç", "ô", "ü"]
Variable label max length	failed (6)	Variable labels should not exceed the defined number of characters (79 characters)

Deployment

Install from UKDS GitHub page

QAMyData

The UKDS website contains a detailed install guide and further materials

Can be partitioned off from other resources (e.g. concerns about disclosive data being uploaded) e.g. A Survey Data Quality Assurance Service Based on QAMyData developed by the Australian Consortium for Social and Political Research Inc.

Live demo…

Statistical Disclosure Control Freeware

- **Amnesia**
  - OpenAire anonymisation tool, with both locally and online run functionalities
  - documentation

- **μ-Argus**
  - standalone software recommended by Eurostat for government statisticians
  - software and manual

- **ARX**
  - comprehensive open source software for anonymizing sensitive personal data
  - software and documentation

- **R tool - sdcMicro (scripting + GUI)**
  - R package (and free dependable software R and RStudio)

Commonly used techniques

- Aggregate categories to reduce precision

  e.g. birth year vs. date of birth, occupational categories, area vs. village name
- Band ages or restrict upper /lower ranges of a variable to disguise outliers

  e.g. incomes, expenditure
- Use standard coding frames – e.g. SOC2010 for employment
- Generalise meaning of detailed text

  e.g. occupational expertise
- Combine variables

  e.g. creating non-disclosive rural / urban variable from place variables

Document any changes made

Published guides: ONS *Disclosure control guidance for microdata produced from social surveys*

Evaluating disclosure risk using tools

- Useful for providing comparison between SDC methods

- Quick and easy to explore what changes have biggest effect

- Allows for reproducibility

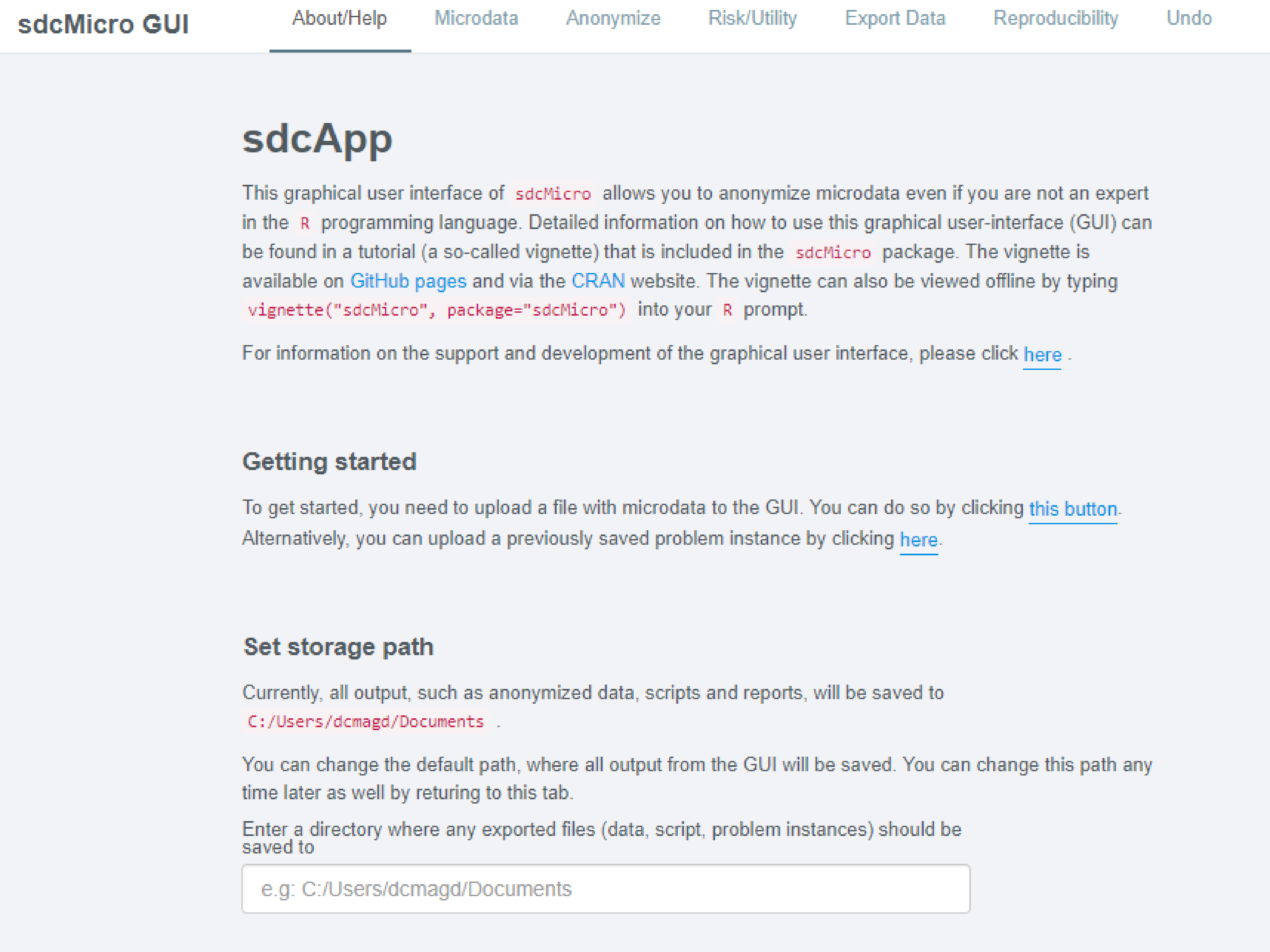
- More problematic when trying to define absolute risk - the numbers might have no real meaning –

- Skill-intensive

sdcMicro

- free, R-based open-source package
- multiple statistical disclosure control methods (perturbative and non-perturbative)
- multiple risk assessment methods (individual risk, global risk)
- locally run friendly GUI (no R knowledge needed)
- permits reproducibility (script and reports)
- well-documented (several online resources)

Live demo…



sdcApp

This graphical user interface of `sdcMicro` allows you to anonymize microdata even if you are not an expert in the `R` programming language. Detailed information on how to use this graphical user-interface (GUI) can be found in a tutorial (a so-called vignette) that is included in the `sdcMicro` package. The vignette is available on GitHub pages and via the CRAN website. The vignette can also be viewed offline by typing `vignette("sdcMicro", package="sdcMicro")` into your `R` prompt.

For information on the support and development of the graphical user interface, please click here .

Getting started

To get started, you need to upload a file with microdata to the GUI. You can do so by clicking this button. Alternatively, you can upload a previously saved problem instance by clicking here.

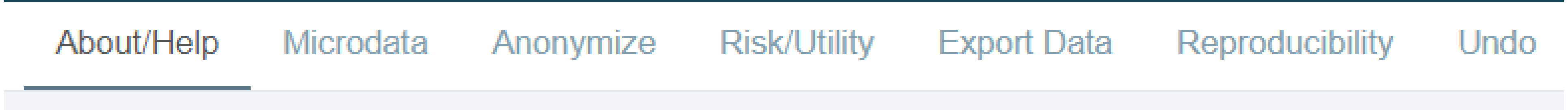
Set storage path

Currently, all output, such as anonymized data, scripts and reports, will be saved to `C:/Users/dcmagd/Documents` .

You can change the default path, where all output from the GUI will be saved. You can change this path any time later as well by returing to this tab.

Enter a directory where any exported files (data, script, problem instances) should be saved to

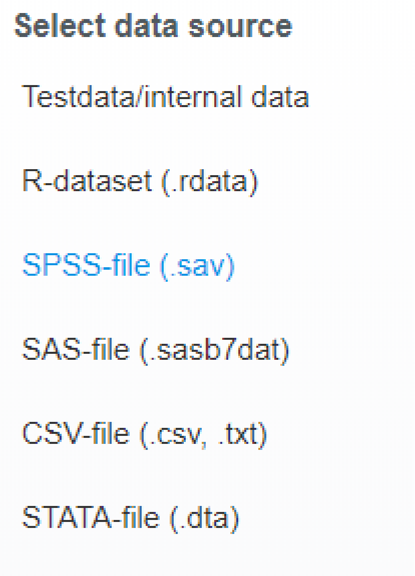
e.g: C:/Users/dcmagd/Documents

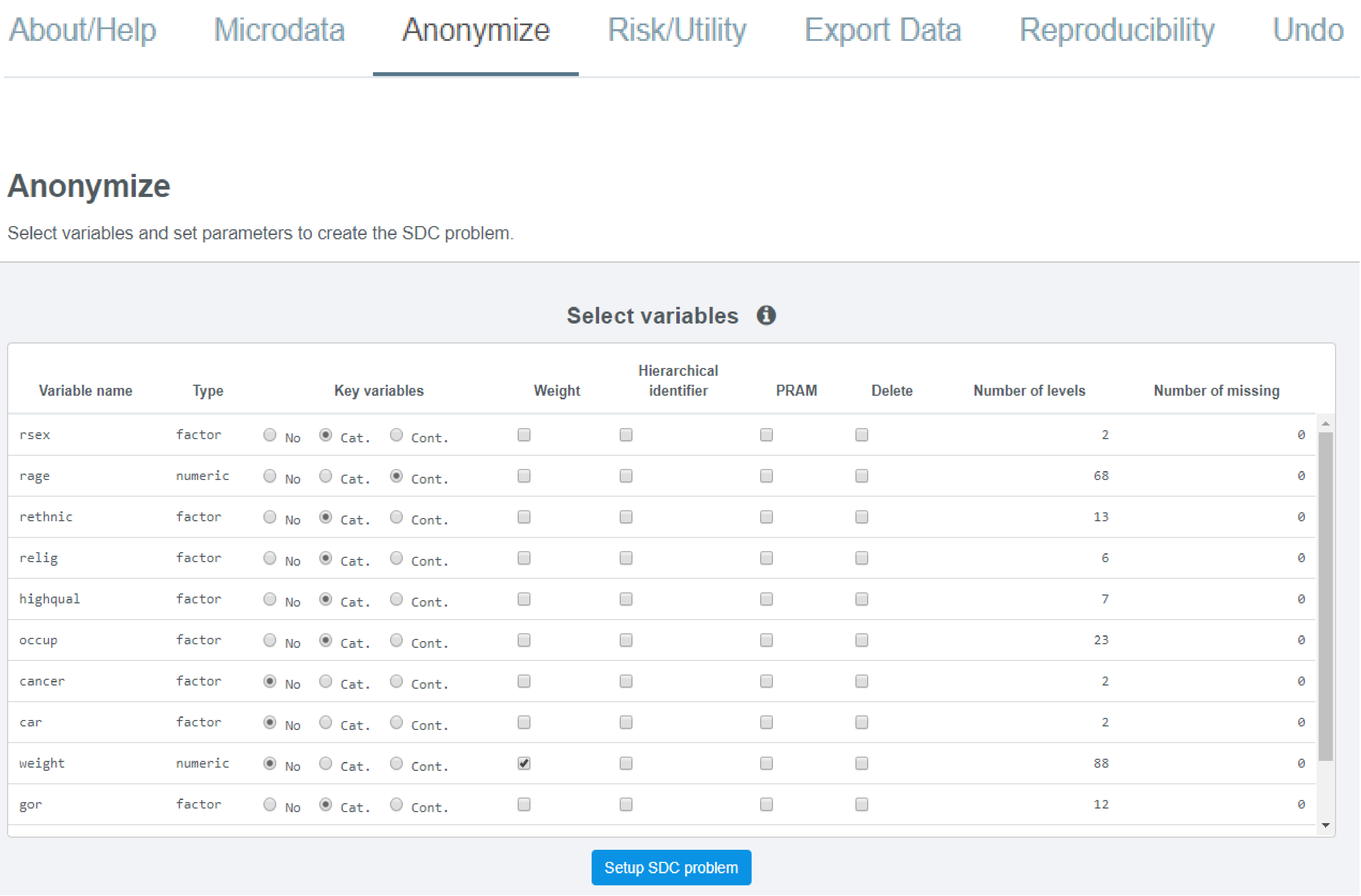


sdcMicro GUI has 7 main menus:

- About/Help
- Microdata
- Anonymize
- Risk/Utility
- Export Data
- Reproducibility
- Undo

All menus have various options displayed on the left-hand side such as:



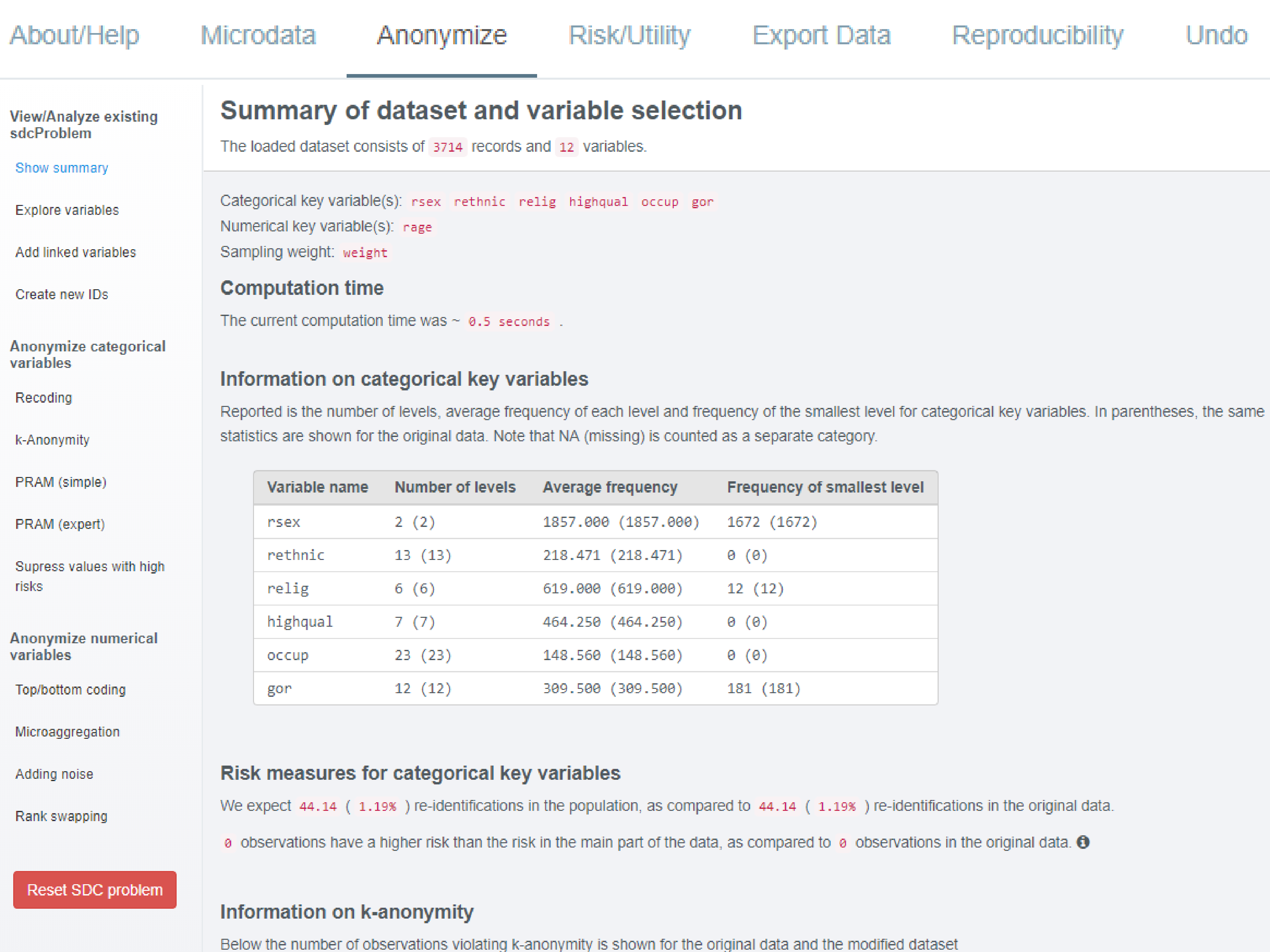


Anonymize

Select variables and set parameters to create the SDC problem.

Select variables

Variable name	Type	Key variables	Weight	Hierarchical identifier	PRAM	Delete	Number of levels	Number of missing
rsex	factor	○ No ● Cat. ○ Cont.	□	□	□	□	2	0
rage	numeric	○ No ○ Cat. ● Cont.	□	□	□	□	68	0
rethnic	factor	○ No ● Cat. ○ Cont.	□	□	□	□	13	0
relig	factor	○ No ● Cat. ○ Cont.	□	□	□	□	6	0
highqual	factor	○ No ● Cat. ○ Cont.	□	□	□	□	7	0
occup	factor	○ No ● Cat. ○ Cont.	□	□	□	□	23	0
cancer	factor	● No ○ Cat. ○ Cont.	□	□	□	□	2	0
car	factor	● No ○ Cat. ○ Cont.	□	□	□	□	2	0
weight	numeric	● No ○ Cat. ○ Cont.	☑	□	□	□	88	0
gor	factor	○ No ● Cat. ○ Cont.	□	□	□	□	12	0



Summary of dataset and variable selection

The loaded dataset consists of 3714 records and 12 variables.

Categorical key variable(s): rsex rethnic relig highqual occup gor
Numerical key variable(s): rage
Sampling weight: weight

Computation time

The current computation time was ~ 0.5 seconds .

Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level
rsex	2 (2)	1857.000 (1857.000)	1672 (1672)
rethnic	13 (13)	218.471 (218.471)	0 (0)
relig	6 (6)	619.000 (619.000)	12 (12)
highqual	7 (7)	464.250 (464.250)	0 (0)
occup	23 (23)	148.560 (148.560)	0 (0)
gor	12 (12)	309.500 (309.500)	181 (181)

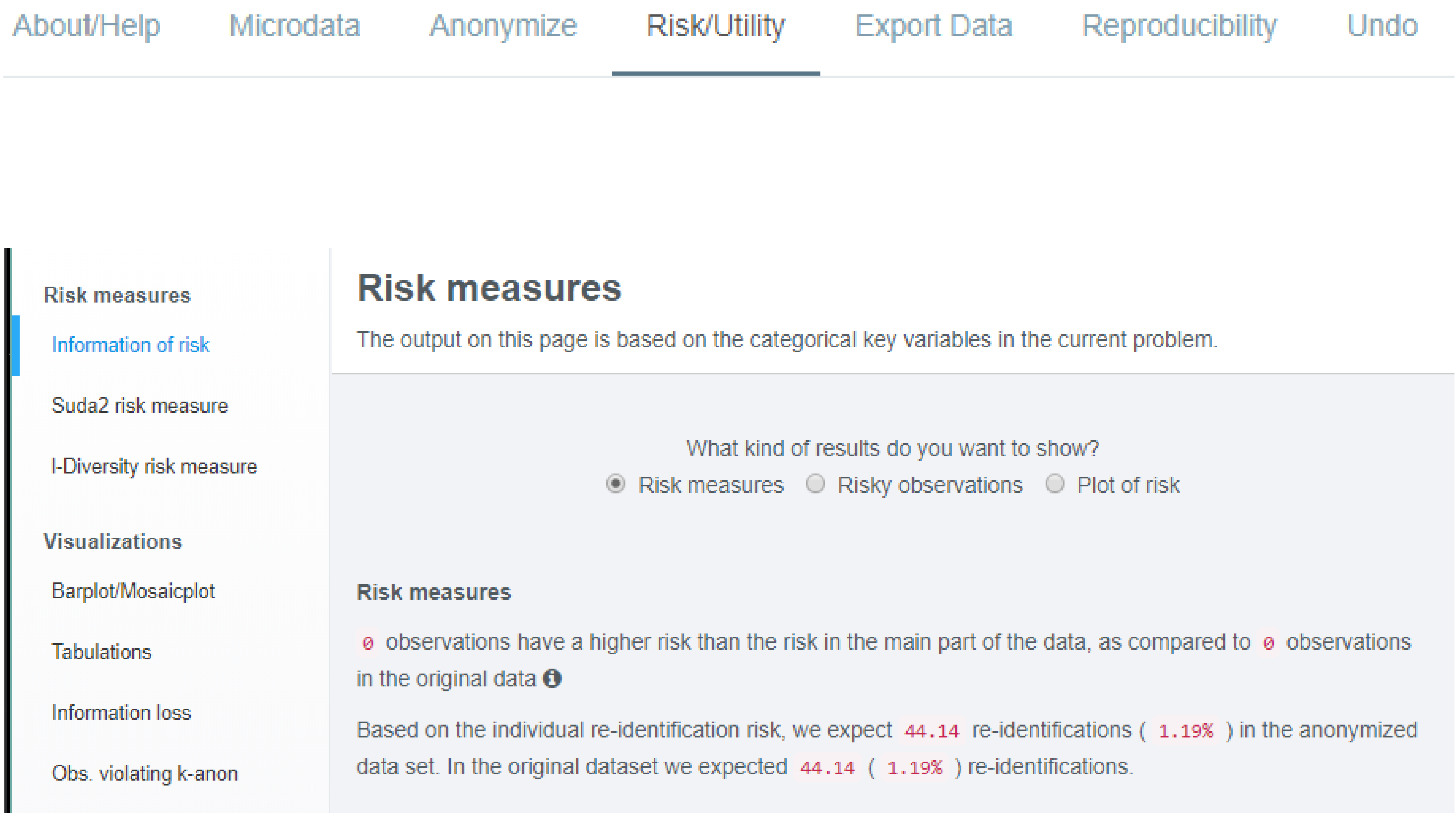
Risk measures for categorical key variables

We expect 44.14 (1.19%) re-identifications in the population, as compared to 44.14 (1.19%) re-identifications in the original data.

0 observations have a higher risk than the risk in the main part of the data, as compared to 0 observations in the original data.

Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset



Risk measures

The output on this page is based on the categorical key variables in the current problem.

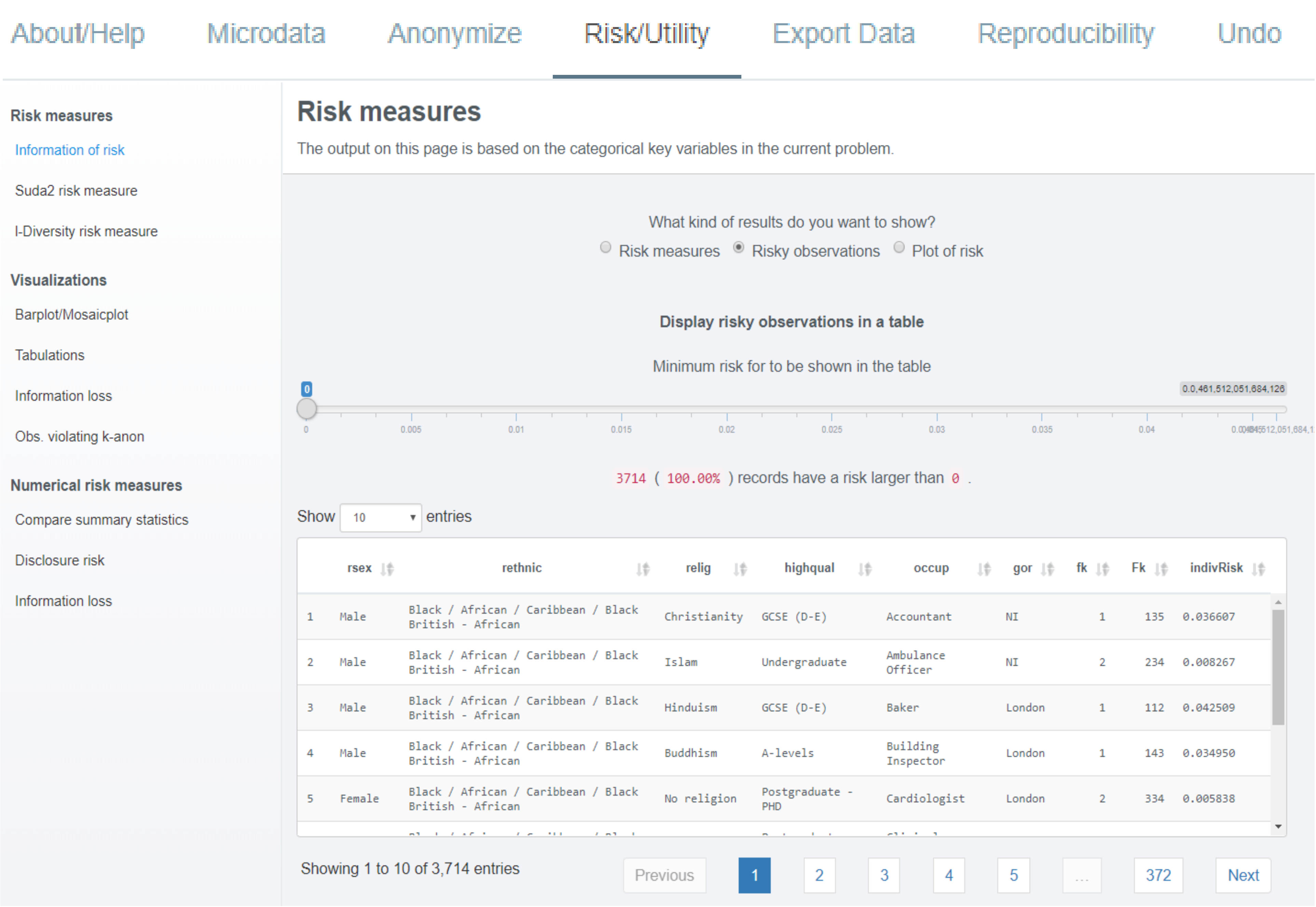
What kind of results do you want to show?

● Risk measures ○ Risky observations ○ Plot of risk

Risk measures

`0` observations have a higher risk than the risk in the main part of the data, as compared to `0` observations in the original data ℹ

Based on the individual re-identification risk, we expect `44.14` re-identifications (`1.19%`) in the anonymized data set. In the original dataset we expected `44.14` (`1.19%`) re-identifications.



Risk measures

The output on this page is based on the categorical key variables in the current problem.

What kind of results do you want to show?

○ Risk measures ● Risky observations ○ Plot of risk

Display risky observations in a table

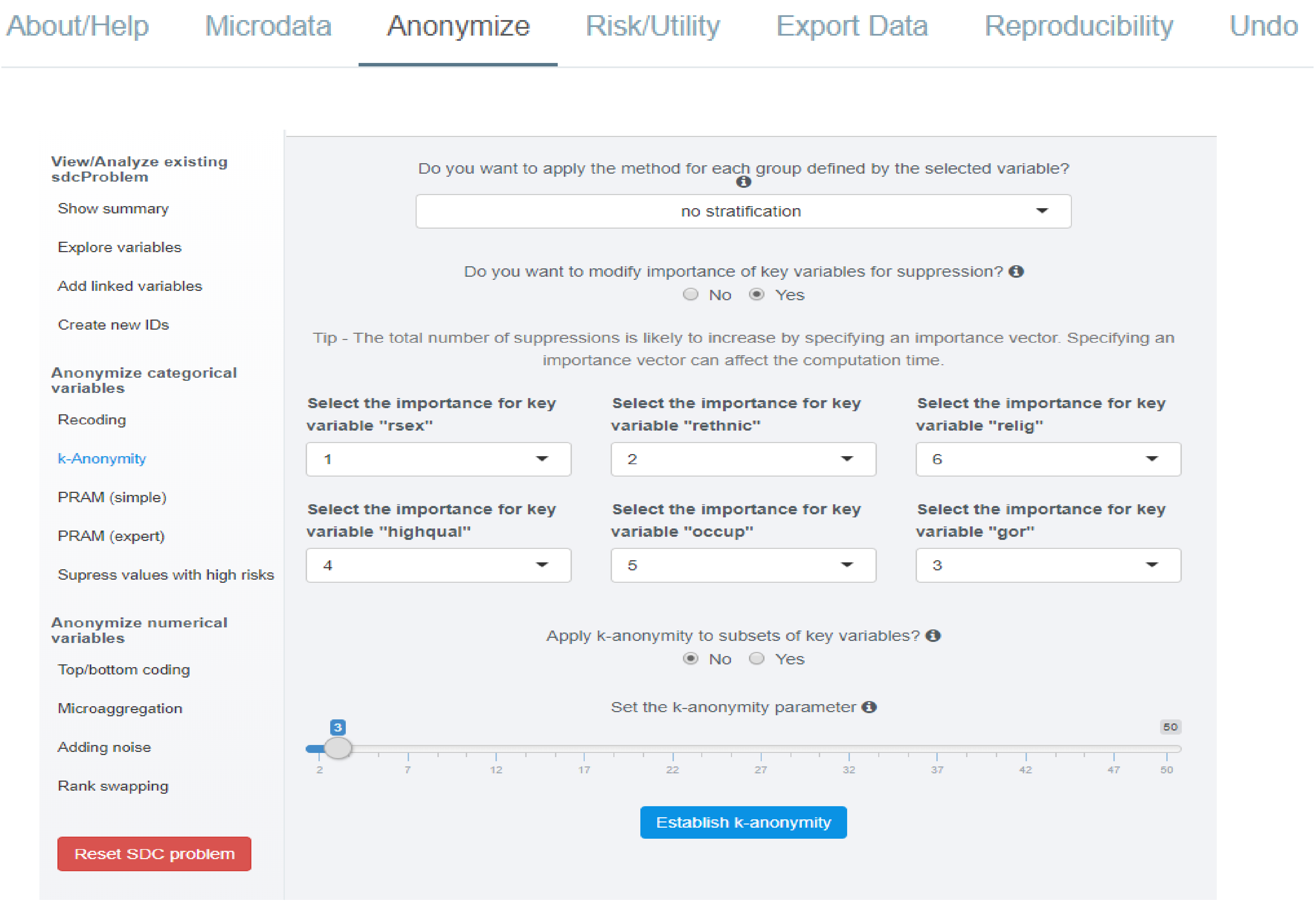
Minimum risk for to be shown in the table

3714 (100.00%) records have a risk larger than 0 .

Show 10 entries

	rsex	rethnic	relig	highqual	occup	gor	fk	Fk	indivRisk
1	Male	Black / African / Caribbean / Black British - African	Christianity	GCSE (D-E)	Accountant	NI	1	135	0.036607
2	Male	Black / African / Caribbean / Black British - African	Islam	Undergraduate	Ambulance Officer	NI	2	234	0.008267
3	Male	Black / African / Caribbean / Black British - African	Hinduism	GCSE (D-E)	Baker	London	1	112	0.042509
4	Male	Black / African / Caribbean / Black British - African	Buddhism	A-levels	Building Inspector	London	1	143	0.034950
5	Female	Black / African / Caribbean / Black British - African	No religion	Postgraduate - PHD	Cardiologist	London	2	334	0.005838

Showing 1 to 10 of 3,714 entries



Do you want to apply the method for each group defined by the selected variable?

no stratification

Do you want to modify importance of key variables for suppression?

○ No ● Yes

Tip - The total number of suppressions is likely to increase by specifying an importance vector. Specifying an importance vector can affect the computation time.

Select the importance for key variable "rsex": 1

Select the importance for key variable "rethnic": 2

Select the importance for key variable "relig": 6

Select the importance for key variable "highqual": 4

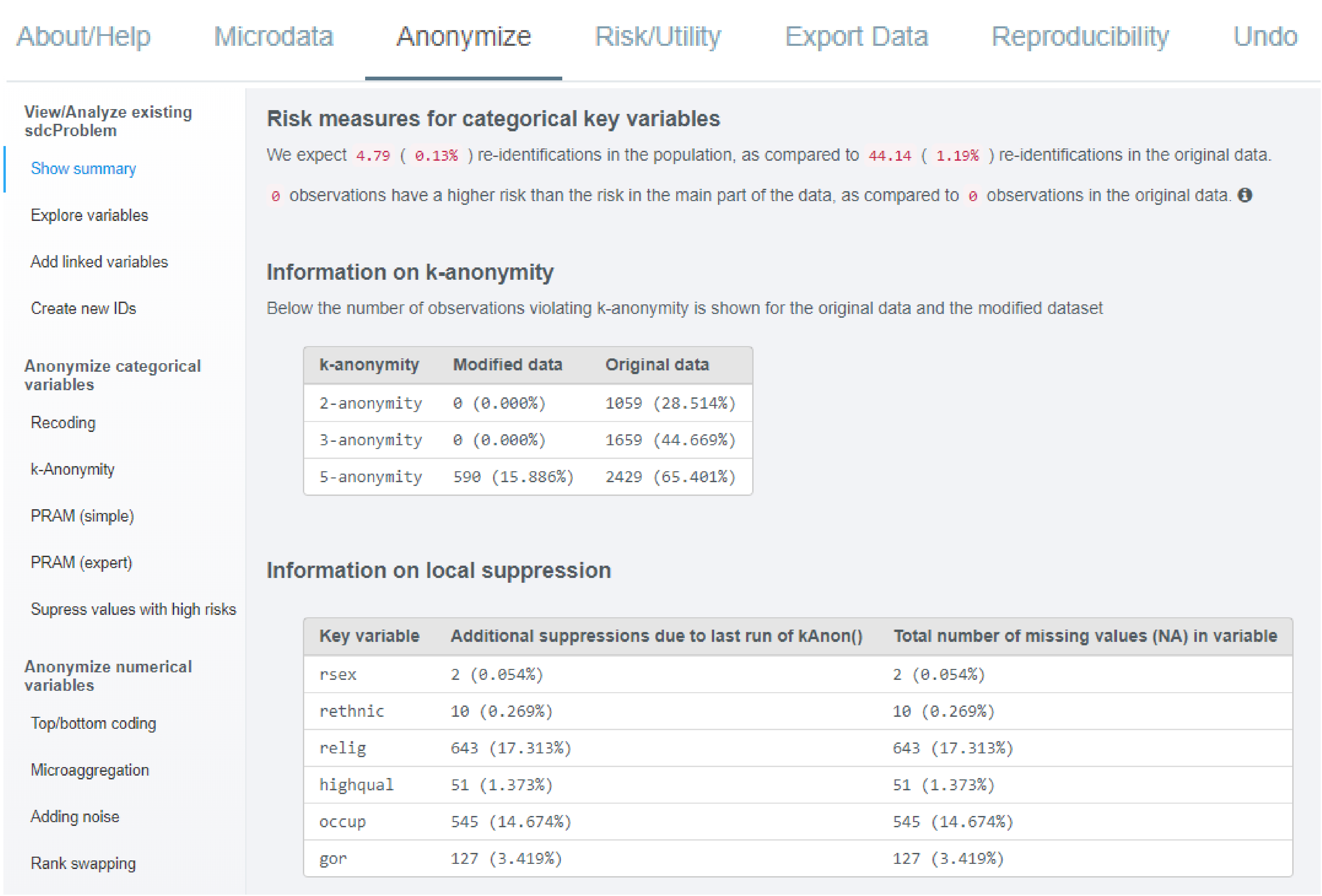
Select the importance for key variable "occup": 5

Select the importance for key variable "gor": 3

Apply k-anonymity to subsets of key variables?

● No ○ Yes

Set the k-anonymity parameter: 3



Risk measures for categorical key variables

We expect `4.79` (`0.13%`) re-identifications in the population, as compared to `44.14` (`1.19%`) re-identifications in the original data.

`0` observations have a higher risk than the risk in the main part of the data, as compared to `0` observations in the original data. ℹ

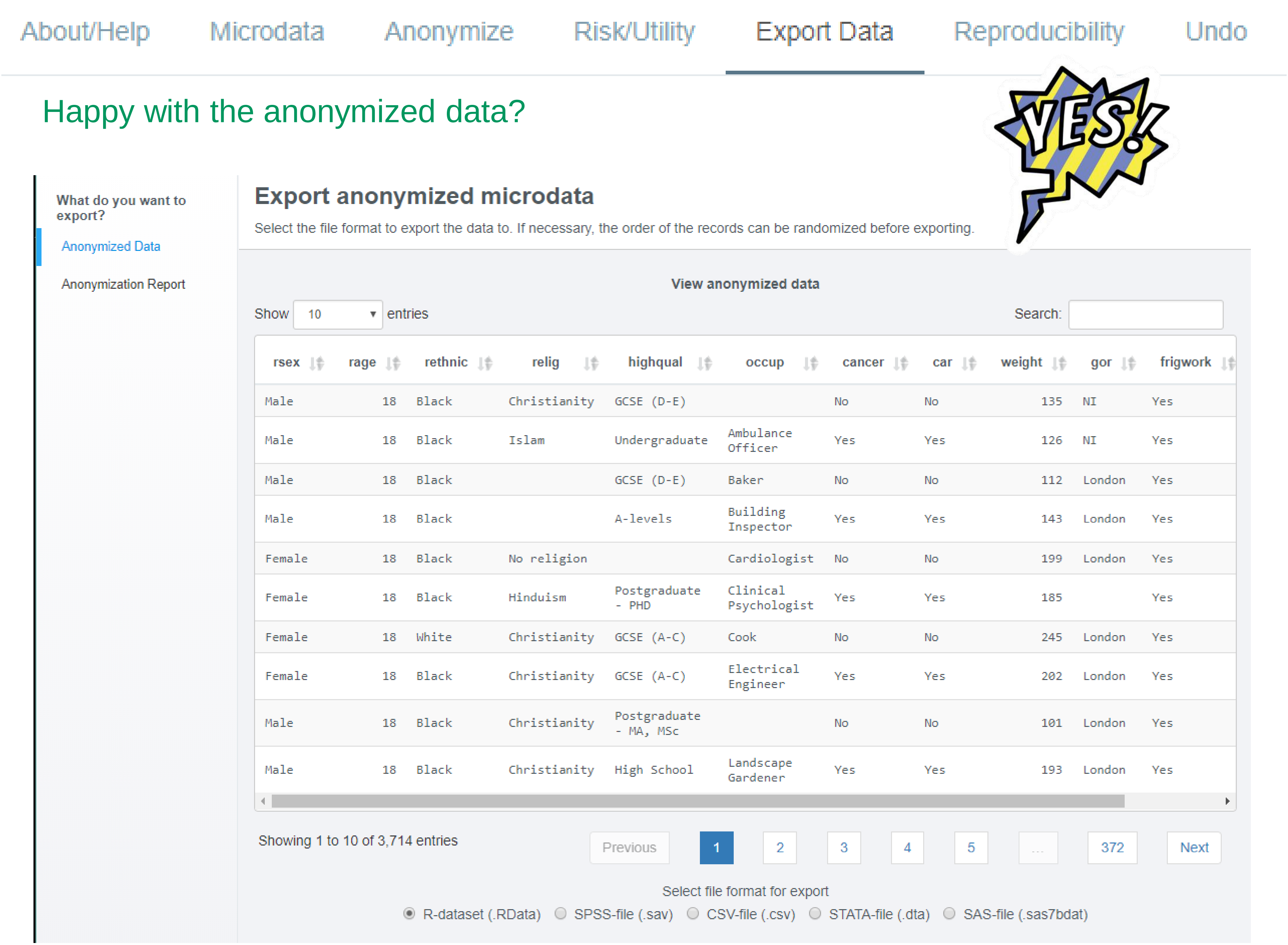
Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	1059 (28.514%)
3-anonymity	0 (0.000%)	1659 (44.669%)
5-anonymity	590 (15.886%)	2429 (65.401%)

Information on local suppression

Key variable	Additional suppressions due to last run of kAnon()	Total number of missing values (NA) in variable
rsex	2 (0.054%)	2 (0.054%)
rethnic	10 (0.269%)	10 (0.269%)
relig	643 (17.313%)	643 (17.313%)
highqual	51 (1.373%)	51 (1.373%)
occup	545 (14.674%)	545 (14.674%)
gor	127 (3.419%)	127 (3.419%)



Happy with the anonymized data?

What do you want to export?

- Anonymized Data
- Anonymization Report

Export anonymized microdata

Select the file format to export the data to. If necessary, the order of the records can be randomized before exporting.

View anonymized data

Show 10 ▾ entries

Search:

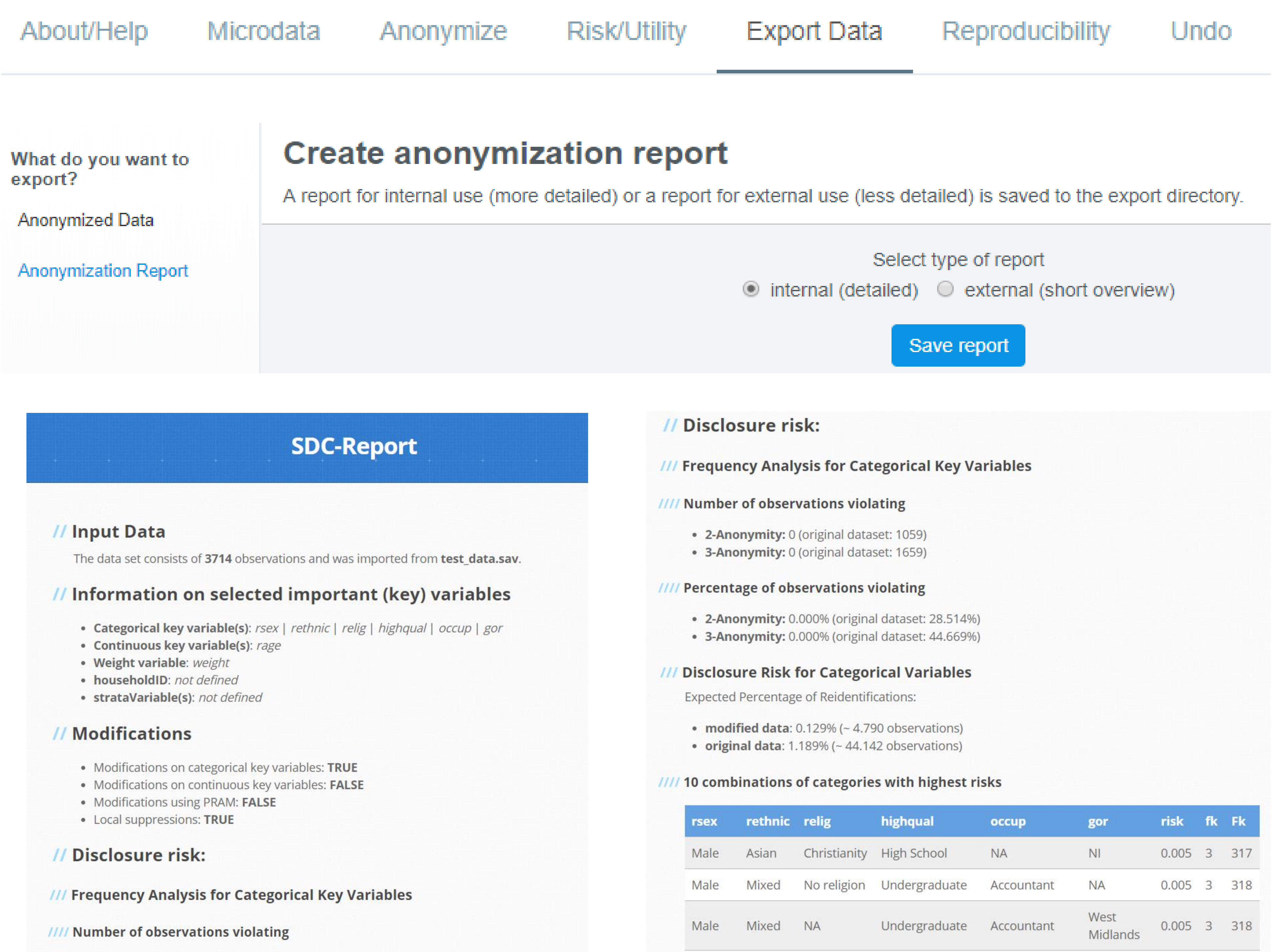
rsex	rage	rethnic	relig	highqual	occup	cancer	car	weight	gor	frigwork
Male	18	Black	Christianity	GCSE (D-E)		No	No	135	NI	Yes
Male	18	Black	Islam	Undergraduate	Ambulance Officer	Yes	Yes	126	NI	Yes
Male	18	Black		GCSE (D-E)	Baker	No	No	112	London	Yes
Male	18	Black		A-levels	Building Inspector	Yes	Yes	143	London	Yes
Female	18	Black	No religion		Cardiologist	No	No	199	London	Yes
Female	18	Black	Hinduism	Postgraduate - PHD	Clinical Psychologist	Yes	Yes	185		Yes
Female	18	White	Christianity	GCSE (A-C)	Cook	No	No	245	London	Yes
Female	18	Black	Christianity	GCSE (A-C)	Electrical Engineer	Yes	Yes	202	London	Yes
Male	18	Black	Christianity	Postgraduate - MA, MSc		No	No	101	London	Yes
Male	18	Black	Christianity	High School	Landscape Gardener	Yes	Yes	193	London	Yes

Showing 1 to 10 of 3,714 entries

Previous 1 2 3 4 5 … 372 Next

Select file format for export

● R-dataset (.RData) ○ SPSS-file (.sav) ○ CSV-file (.csv) ○ STATA-file (.dta) ○ SAS-file (.sas7bdat)



What do you want to export?

Create anonymization report

A report for internal use (more detailed) or a report for external use (less detailed) is saved to the export directory.

Select type of report

● internal (detailed) ○ external (short overview)

SDC-Report

// Input Data

The data set consists of **3714** observations and was imported from **test_data.sav**.

// Information on selected important (key) variables

- **Categorical key variable(s)**: *rsex* | *rethnic* | *relig* | *highqual* | *occup* | *gor*
- **Continuous key variable(s)**: *rage*
- **Weight variable**: *weight*
- **householdID**: *not defined*
- **strataVariable(s)**: *not defined*

// Modifications

- Modifications on categorical key variables: **TRUE**
- Modifications on continuous key variables: **FALSE**
- Modifications using PRAM: **FALSE**
- Local suppressions: **TRUE**

// Disclosure risk:

/// Frequency Analysis for Categorical Key Variables

//// Number of observations violating

// Disclosure risk:

/// Frequency Analysis for Categorical Key Variables

//// Number of observations violating

- **2-Anonymity:** 0 (original dataset: 1059)
- **3-Anonymity:** 0 (original dataset: 1659)

//// Percentage of observations violating

- **2-Anonymity:** 0.000% (original dataset: 28.514%)
- **3-Anonymity:** 0.000% (original dataset: 44.669%)

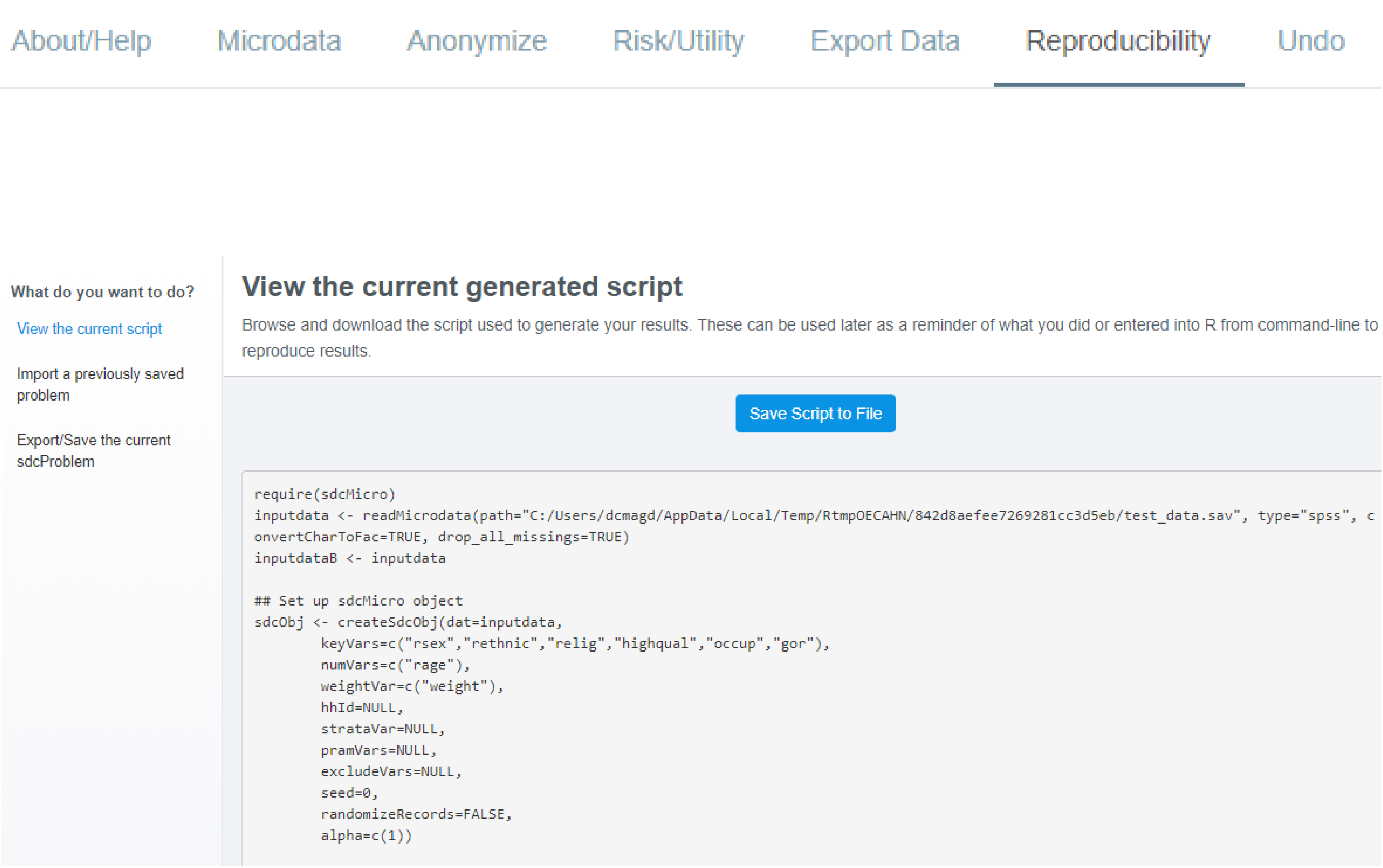
/// Disclosure Risk for Categorical Variables

Expected Percentage of Reidentifications:

- **modified data**: 0.129% (~ 4.790 observations)
- **original data**: 1.189% (~ 44.142 observations)

//// 10 combinations of categories with highest risks

rsex	rethnic	relig	highqual	occup	gor	risk	fk	Fk
Male	Asian	Christianity	High School	NA	NI	0.005	3	317
Male	Mixed	No religion	Undergraduate	Accountant	NA	0.005	3	318
Male	Mixed	NA	Undergraduate	Accountant	West Midlands	0.005	3	318



View the current generated script

Browse and download the script used to generate your results. These can be used later as a reminder of what you did or entered into R from command-line to reproduce results.

```
require(sdcMicro)
inputdata <- readMicrodata(path="C:/Users/dcmagd/AppData/Local/Temp/RtmpOECAHN/842d8aefee7269281cc3d5eb/test_data.sav", type="spss", c
onvertCharToFac=TRUE, drop_all_missings=TRUE)
inputdataB <- inputdata

## Set up sdcMicro object
sdcObj <- createSdcObj(dat=inputdata,
        keyVars=c("rsex","rethnic","relig","highqual","occup","gor"),
        numVars=c("rage"),
        weightVar=c("weight"),
        hhId=NULL,
        strataVar=NULL,
        pramVars=NULL,
        excludeVars=NULL,
        seed=0,
        randomizeRecords=FALSE,
        alpha=c(1))
```

sdcMicro further resources

- sdcMicro Reference Manual
- sdcMicro GUI Vignette
- Guidelines for statistical disclosure control using sdcMicro Vignette
- sdcMicro Git Page
- Matthias Templ, Alexander Kowarik, Bernhard Meindl (2015). *Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro*. Journal of Statistical Software, 67(4), 1-36. doi:10.18637/jss.v067.i04

Questions

Cristina Magder

dcmagd@essex.ac.uk

UKDS Data Sharing Queries

datasharing@ukdataservice.ac.uk

QAMyData mailbox

qamydata@ukdataservice.ac.uk

UK Data Service