Reducing risk: An introduction to data anonymization

Safe Sharing: dream or myth?

Presented by Kristi Thompson

Western
UNIVERSITY · CANADA

Some context

- Canada's Tri-Agency: Grant recipients are required to deposit data.
- NIH: Encourages the sharing of data whenever possible.
- NSF: investigators are expected to share with other researchers…
- Many journals: Show us your data.
- This does not mean sharing all data openly. But most current academic and journal repositories are designed to make open sharing the most convenient option.
- I'm approaching this as a research data management librarian – someone expected to help researchers comply with these policies.
  - Are researchers expected to understand when data deidentification is needed to comply with data sharing mandates? Are data curators? Research ethics boards / IRBs?

Background and key concepts

IDENTIFIERS, QUASI-IDENTIFIERS, RISK

Direct Identifiers

- Any information collected by the researcher that places study participants at immediate risk of being reidentified
- Full or parts of: Names, addresses, telephone numbers, or any identifiers used by the researchers to link data to one of the above
- Detailed geography (areas containing less than 20,000 people is a rule of thumb - HIPAA)
- IP addresses and other information that may be associated with a computer
- Exact dates linked to individuals or events are highly identifying
- HIPAA recognizes 18 personal identifiers that will qualify data as personal health information; the BMJ compiled a list of 28 based on multiple international research guidelines. There's a Canadian list in the text Research Data Management in the Canadian Context.

Quasi-identifiers

- Characteristics relating to individuals that could be linked with other data sources to violate the confidentiality of individuals
- A variable should be considered a quasi-identifier if an attacker could plausibly match that variable to information from another source to determine the identity of an individual
- Some variables may be used in combination to derive quasi-identifiers, e.g. community size (at first glance not particularly identifying) could be combined with a broader geographic grouping to infer location more precisely

Hidden identifiers

- Quasi-identifiers are commonly thought of as demographic variables and socio-economic variables that have the potential to be linked with other data sources to violate the confidentiality of participants, or to be recognized by a person acquainted with the survey respondent.
  - Specific examples include age, gender identity, income, occupation, industry / place of work, geography, ethnic and immigration variables
- Potentially, membership in specific organizations, use of specific services
- Variables that relate to geography in any way need to be treated with extreme caution
  - Potential community identifiers can include features like presence of a university hospital or international airport
  - E.G. variable giving distance to nearest emergency department
  - Need to be considered alongside any contextual information about the dataset

Risk – a technical definition

- Risk is created when:
  - Variables can isolate individuals in the dataset
  - Identifying information can be matched to persistent information that an attacker may reasonably have access to
- A set of records that has the same values on all quasi-identifiers is called an *equivalence class*
- An equivalence class of one corresponds to an individual who is unique in the dataset on some combination of characteristics. Such a person may be at risk of being identified.
  - This person is called a *sample unique*. If your survey is a complete sample of some population, this person is also a *population unique*.

Assessing and dealing with risk: statistical disclosure risk assessment

an introduction to

K-anonymity

Assessing quasi-identifiers – first pass

- Quasi-identifying variables containing groups with small numbers of respondents (e.g. a religion variable with 3 individual responses of "Buddhism") pose high risk.
- Extreme values (more than 10 children; very high income) pose high risk
- Size of identifiable groups *in the general population* also need to be considered
  - There may be only one person from Winnipeg in your random digit cell phone user survey, but if your survey doesn't narrow it down any further than that, that person is pretty safe
- Contextual information that accompanies the data should also be part of the analysis
  - If it is clear from the context of your research that all your interview subjects worked at a particular tool and die plant in Oshawa, that narrows things down quite a bit

Common sense (can only take you so far)

- Look at the demographic variables in the dataset and consider describing an individual to a friend using only the values of those variables.
- “I’m thinking of a person living in Toronto who is female, married, has a University degree, is between the ages of 40 and 55 and has an income of between 60 and 75 thousand dollars.”
  - Even if there is only one such person in the dataset, this is not enough information to create risk…
  - **UNLESS** you know this is a survey of soccer referees …
- Also, consider unusual combinations of variables – let’s say someone belongs to the under-17 age group and responded that they were married.
- How do you figure this out without needing to know every single combination in the data?

K-anonymity

- K-anonymity is a mathematical approach to demonstrating that a dataset is anonymized
- Concept: it should not be possible to isolate fewer than K individual cases in your dataset based on any combination of identifying variables
- That is, a record cannot be distinguished from K-1 other records in its equivalence class.
- K is a number set by the researcher; three and five are both commonly used

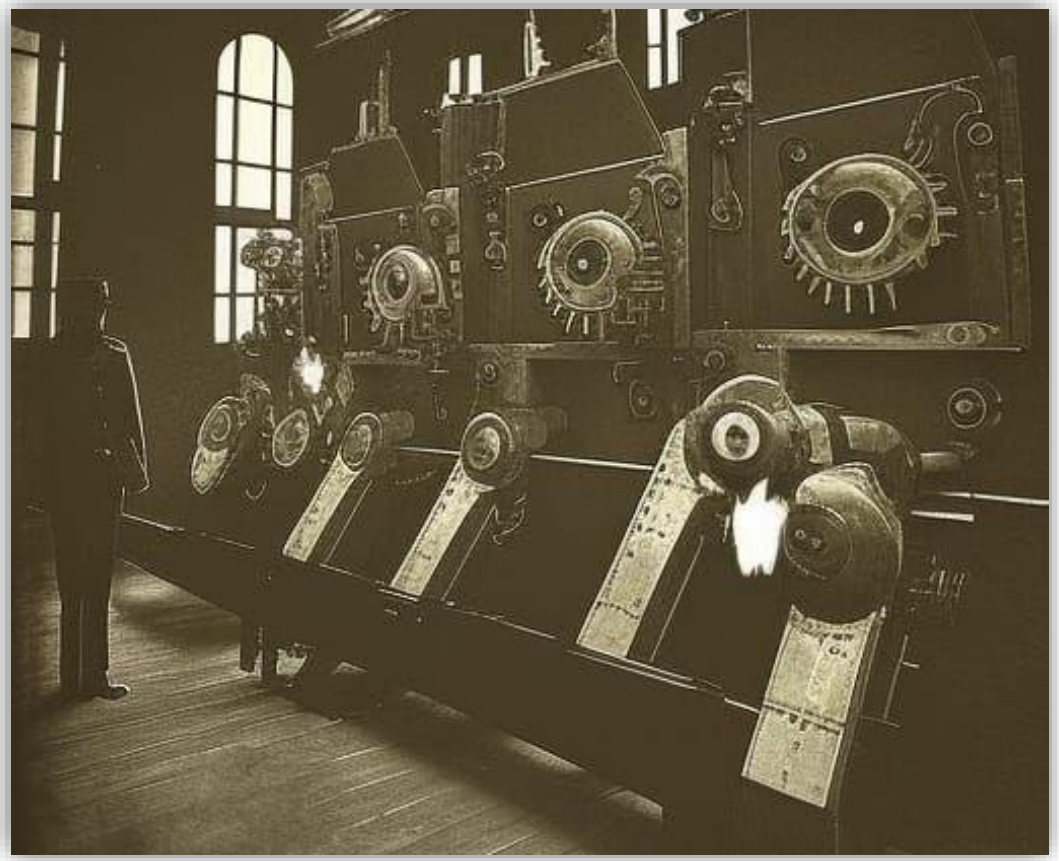
Equivalence classes and “data twins”

- It should not be possible to isolate fewer than k individual cases in your dataset based on any combination of identifying variables
- Cases 1, 6 and 13 form an equivalence class with k=3
  - Each case in the equivalence class has 2 “data twins”
- Case 14 has no data twins – it is a sample unique
- A dataset’s k is the size of the smallest equivalence class in the dataset – in this case 1.

ID	Gender	AgeGrp	EthnicGrp
1	M	25-30	1
2	F	16-24	1
3	M	25-30	2
4	M	16-24	1
5	F	31-45	1
6	M	25-30	1
7	F	16-24	1
8	F	31-45	1
9	F	31-45	2
10	M	25-30	2
11	M	16-24	1
12	F	25-30	1
13	M	25-30	1
14	F	16-24	2
15	F	31-45	1

Data reduction – global reduction and local suppression

- Global data reduction
  - Grouping into categories e.g. age in 10 year increments
  - For already categorical variables, merging into larger groups
  - Complete removal of risky variables from the dataset
- Local suppression
  - Deleting individual cases or responses
  - For example, a member of the ‘under 17’ age group who responded ‘married’ might have their response deleted



Checking k-anonymity

- Stata statistical language:

```
egen equivalence_group= group(var1 var2 var3 var4 var5)
* create a variable to count cases in each equivalence group
sort equivalence_group
by equivalence_group: gen equivalence_size =_N
tab equivalence_group if equivalence_size < 5, sort
```

- R statistical language

```
library('plyr’)
# Figure out what equivalence classes there are, and how many cases in each
equivalence class.
dfunique <- ddply(df, .(var1, var2, var3, var4, var5), nrow)
dfunique <- dfunique[order(dfunique$V1),]
View(dfunique)
```

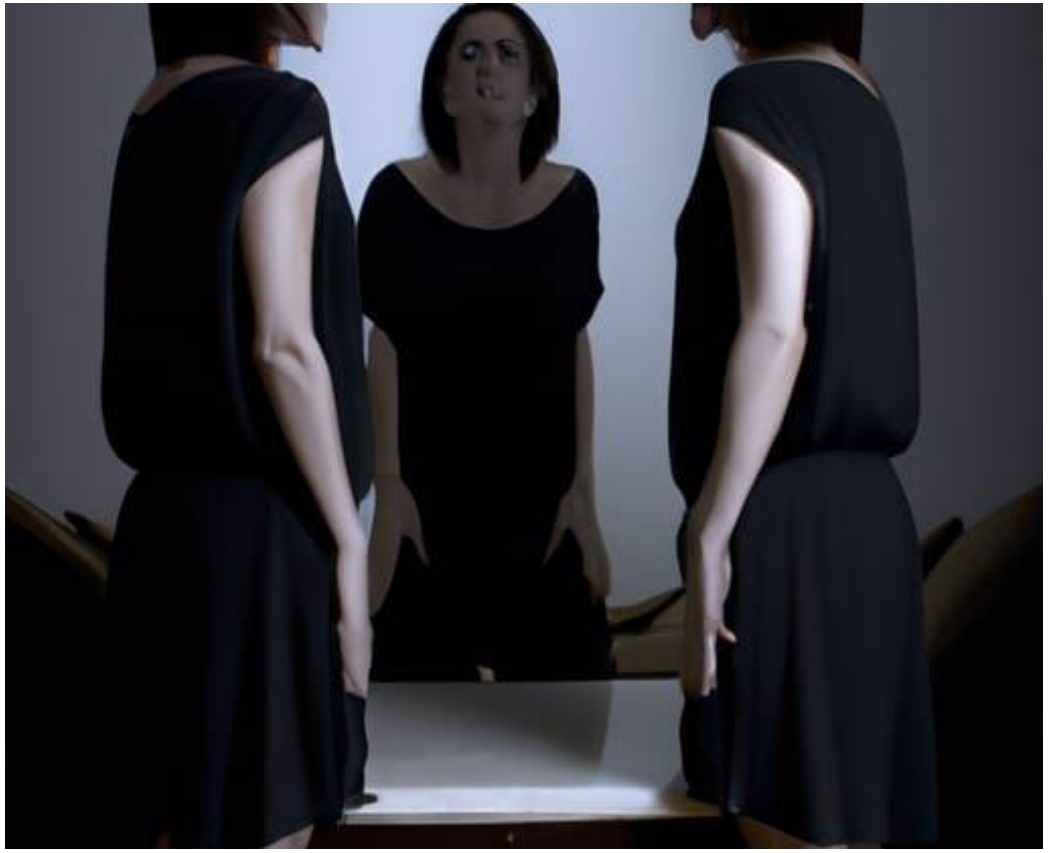
The UK Anonymisation Network Anonymization Decision-Making Framework, appendix B has code for doing this in SPSS.

~~Guaranteed~~ data anonymization

- k-anonymity is intended to be a form of guaranteed data anonymization and is often described as such.
- It guarantees that every record in the anonymized data will be indistinguishable from k-1 other records in the same dataset.

However…

- Research participants are not generally told that no one will know which line of the data file holds their confidential information. They are told their answers to research questions will be kept confidential.



Attribute disclosure

Introducing l-diversity and friends

Attribute Disclosure

- Cases 1, 6 and 13 still form an equivalence class with k=3. So even if you know which people in this survey population match those characteristics, you can't tell which person matches which case

BUT

- They all answered a particular question (about whether their workplace should unionize) the same way
- You now know how all three of them answered this question. Confidentiality had been violated.

ID	Gender	AgeGrp	EthnicGrp	Unionize
1	M	25-30	1	**Y**
2	F	16-24	1	N
3	M	25-30	2	N
4	M	16-24	1	Y
5	F	31-45	1	Y
6	M	25-30	1	**Y**
7	F	16-24	1	N
8	F	31-45	1	Y
9	F	31-45	2	Y
10	M	25-30	2	N
11	M	16-24	1	Y
12	F	25-30	1	Y
13	M	25-30	1	**Y**
14	F	16-24	2	N
15	F	31-45	1	Y

ℓ-diversity and friends

- Extensions of *k*-anonymity, including *p*-anonymity and ℓ-diversity, have been proposed to deal with attribute disclosure; they all involve rules around what values the attributes within an equivalence class should have
- Example: one of the simpler variants, called distinct ℓ-diversity
  - A dataset satisfies distinct ℓ-diversity if, for each group of records in an equivalence class (matching on all their quasi-identifiers) there are at least ℓ different responses for each confidential variable
  - So for our workplace survey, every group of data twins would have to contain both yes and no answers to the "unionize" question, since two would be the maximum possible value for ℓ for this question
  - And this would have to be true for some value of ℓ for every confidential answer in the dataset
- Imagine a typical survey dataset with dozens of questions, each of which needs to be considered for ℓ -diversity for each equivalence class…

Issues with techniques like ℓ-diversity

- Only practical to implement in datasets with very few variables
- No computationally efficient ways of doing these; far too time consuming to be done by hand
  - For some of the more esoteric methods, no theoretical implementations have even been described
  - It's been demonstrated that even in relatively simple cases (such as ℓ-diversity with few attributes) automatedly solving for optimal data utility while protecting privacy is NP hard – meaning, essentially, that the time taken to run such an algorithm increases exponentially with the size of the dataset
- Even if they could be implemented, in most cases achieving anything like distinct ℓ-diversity (or *t*-closeness, or *p*-diversity) would completely destroy the reanalysis value of the dataset, making going to this level of effort to make data shareable rather pointless

The role of sampling

A 50% sample

Surveyed						Not Surveyed			
ID	Gender	AgeGrp	EthnicGrp	Unionize		Gender	AgeGrp	EthnicGrp	Unionize
1	M	25-30	1	**Y**		M	25-30	1	**?**
2	F	16-24	1	N		M	25-30	1	**?**
3	M	25-30	2	N		M	25-30	1	**?**
4	M	16-24	1	Y		F	16-24	1	?
5	F	31-45	1	Y		F	16-24	1	?
6	M	25-30	1	**Y**		M	16-24	2	?
7	F	16-24	1	N		F	31-45	1	?
8	F	31-45	1	Y		M	25-30	1	**?**
9	F	31-45	2	Y		M	25-30	1	**?**
10	M	25-30	2	N		M	31-45	1	?
11	M	16-24	1	Y		F	31-45	1	?
12	F	25-30	1	Y		M	25-30	2	?
13	M	25-30	1	**Y**		M	16-24	1	?
14	F	16-24	2	N		F	31-45	1	?
15	F	31-45	1	Y		F	16-24	2	?

Sampling

- Creates uncertainty that any given individual is in the dataset at all
- A sample unique may not be a population unique
  - Still a concern…
- That is, *if* an equivalence class in the dataset can be assumed to have co-equivalents (data twins) outside the dataset whose opinions or attributes are unknown, *then* attributes are not disclosed by membership in an equivalence class
- This is a reasonable assumption in cases where:
  - k-anonymity is met for k >=5
  - Sample is a small subset of the population it is drawn from
  - There is variation in the attributes being looked at
- Attribute disclosure in the absence of identity disclosure ceases to be a concern in the case of a small sample drawn from a large population, given appropriate levels of variation in the attributes.

Bad examples

And how I got involved with this stuff in the first place

Rescuing messy data

- First became seriously involved with data anonymization due to a data rescue project
- Series of government department datasets released due to an open government mandate
- Versions initially made available were unusable due to missing documentation and general incomprehensibility; documented versions made available on request had not been anonymized.
- Our contact recognized that this was a problem but had no de-identified version of the survey, or resources for fixing it

The first test survey

- Survey of adolescents asking about an ad campaign
- ~1500 respondents, limited demographics, various non-identifying variables
- Five quasi-identifier variables of concern: age (3 categories), sex (2), geographic region (7), visible minority status (2) and indigenous* status (2)
  - 126 Possible equivalence classes (not 168 because visible minority and indigenous status are mutually exclusive as defined (...ask them))
- If these were distributed equally across the dataset, we would expect each equivalence class to contain about 12 cases
- For most real-world variables, some groups will be much larger than others. In practice we had 21 equivalence classes with only a single member, and a total of 42 equivalence classes with less than 5 members

k-anonymity is hard

- Only five quasi-identifier variables, only a few categories each
- Was not able to produce a dataset that satisfied k-anonymity, let alone any more stringent criteria such as l-diversity, while retaining all five variables

The role of sampling, redux

- How risky would it have been to retain the region variable? Were the sample unique cases (the 21 equivalence classes with only a single member) also population uniques?
- Checked by downloading a Census of Canada public use file to produce a dataset that matched my survey but represented the population aged 13-15 in Canada at that time as a whole
- Each sample unique in the survey is estimatd to have a minimum of 369 data twins in the general population – k-anonymity overestimated reidentification risk by a factor of 370!

Second test survey

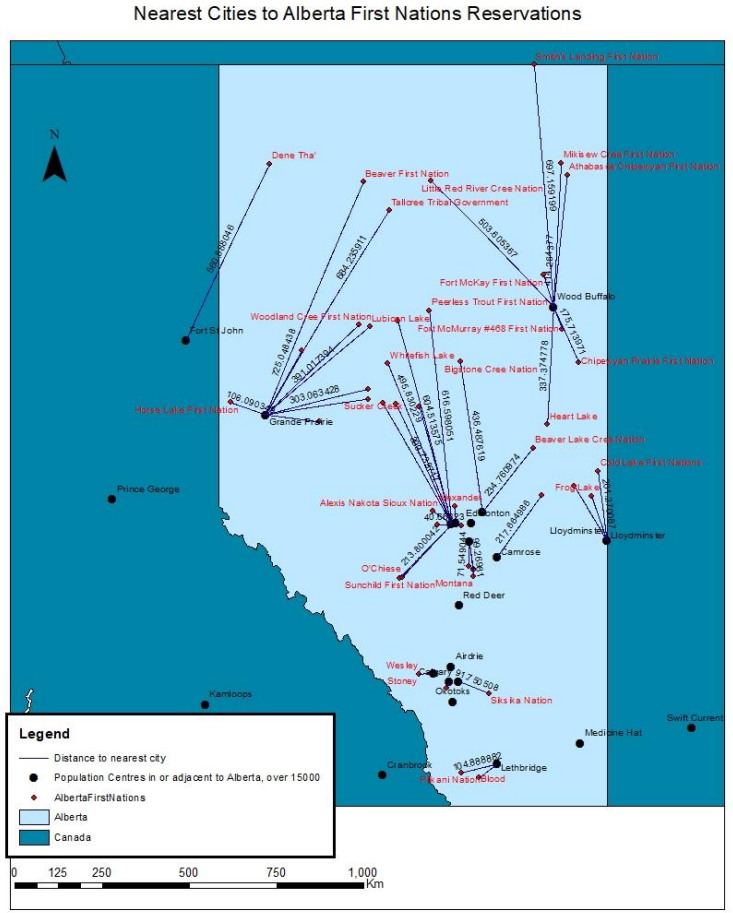
- Additional survey with few demographic variables, but additional data associated with participant location
- Service survey of a population living in small and remote communities. Did not have exact location, community names, or anything obvious like that except province…
- But **did** identify some respondents as being located on a reservation, and also had a variable giving approximate distance to nearest major city
- Original dataset also had partial postal codes that could be used to check guesses

Penetration testing and data linkage

- Means of assessing resistance of de-identified dataset to reidentification of survey participants or their attributes
- Remember: quasi-identifiers contain information that can be matched to persistent information that an attacker may reasonably have access to
- From publicly available information, a data intruder can easily construct a table of reservations by province and their distance from the nearest city
- For each participant, their province of residence and distance from the nearest city can be compared to the entries in the table of reservations by province and distance to the nearest city

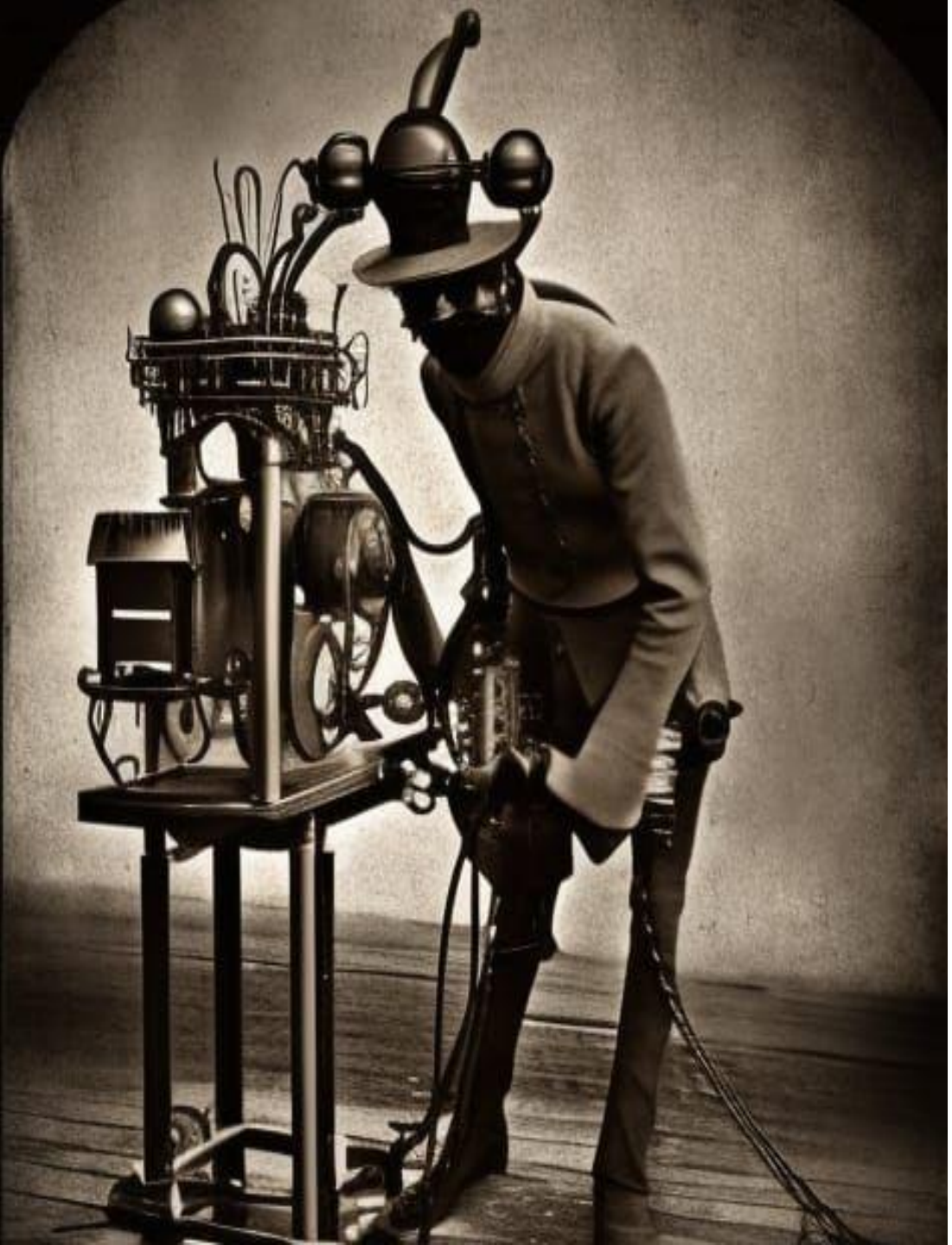
Use of data linkage to construct lists of candidate locations for survey participants

- Hypothetical example:
  - Participant living in Alberta on a reservation
  - Response: 80 km from nearest city
  - 2 possibilities only 10 km apart from one another: Samson and Ermineskin Tribe
- Of over 1,000 individuals surveyed, a single location for their potential place of residence was found for 98
- Of the 98, the (suppressed) value for forward sortation area (first three digits of postal code) was correct for 24 cases (~25% of guesses)
- Accuracy could be improved with access to more specialized GIS tools



Hidden identifiers

- “Distance from respondent’s community to nearest large city” does not generally show up on lists of possible identifiers or quasi-identifiers to check for
  - Community name, yes.
- Variables may be used in combination to derive other quasi-identifiers
  - So burn it all down?
- Anything relating to geography needs to be considered. Similar variables can indirectly identify other groups, such as clubs, organizations, or employers.



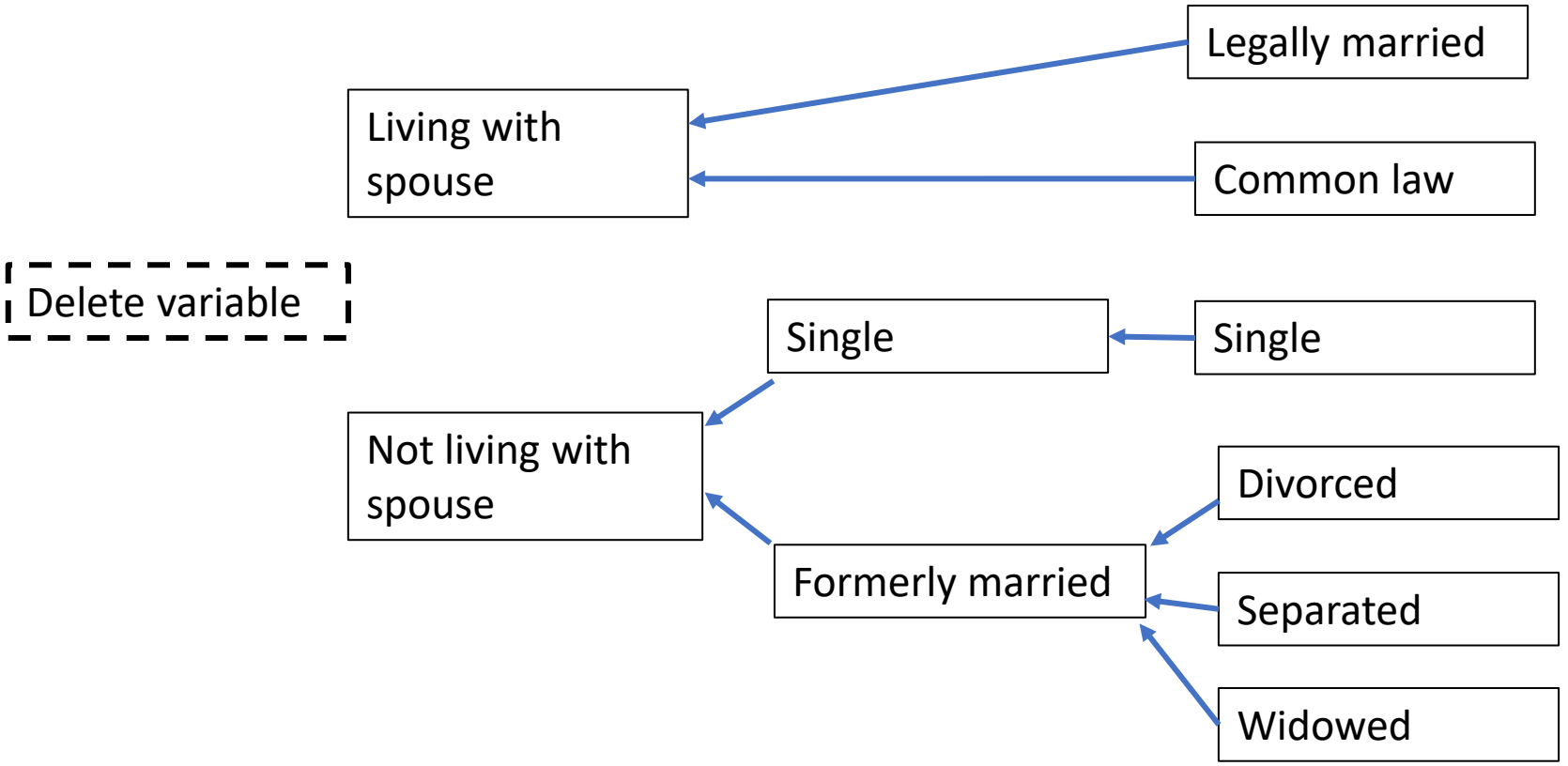
Automation

ALGORITHMIC OPTIONS

Anonymization hierarchies

- Software tools for de-identifying quantitative data that I have investigated take a hierarchy approach to automatically deidentifying data.
- This basically means that the user needs to pre-define possible generalizations for the quasi-identifiers in the dataset, and the program searches for possible solutions and recommends a set of the generalizations to use.
- For datasets with many quasi-identifiers, or cases where several datasets with similar quasi-identifiers need to be deidentified, this might be a useful approach.

Possible hierarchy for the variable “Marital Status”



Tools

- While working on the initial deidentification project and later while contributing to some documents for a working group, I tested several free / open anonymization packages that I found recommended on various lists.
- The one I've used most is SDCMicro, a set of R routines, which I'll briefly discuss. ARX is another freely available, open-source tool that will be discussed in tomorrow's workshop
- Commercial software and for-fee services exist. I have not had the opportunity to try any.

What about AI? Fuzzy Logic?

- Real issue may be Big Data in all its forms. Expands our concept of possible quasi-identifiers
  - Think about databases that track cookies. Could they infer a profile of
    - Locations you've visited?
    - Medicines you've bought?
    - Political opinions?
- AI is not magic. Humans are also intelligent...
- Assuming a master database with everything that could possibly be inferred about your subjects already exists… does the data you are considering sharing make things worse?

Final observations

- Guaranteeing that data has been ‘reasonably’ anonymized is difficult, and the difficulty increases exponentially with the number of variables present.
- k-anonymity can be calculated easily using standard statistical software. Achieving k-anonymity can require a great deal of data modification or suppression, though the role of sampling somewhat mitigates this.
- A dataset that is a complete sample of a small, known population is very difficult to deidentify unless the number of demographic and attribute variables is trivial.
- Software aimed at the general academic survey researcher should not assume special knowledge in the field of data de-identification.
- Big data is scary.

Paws…

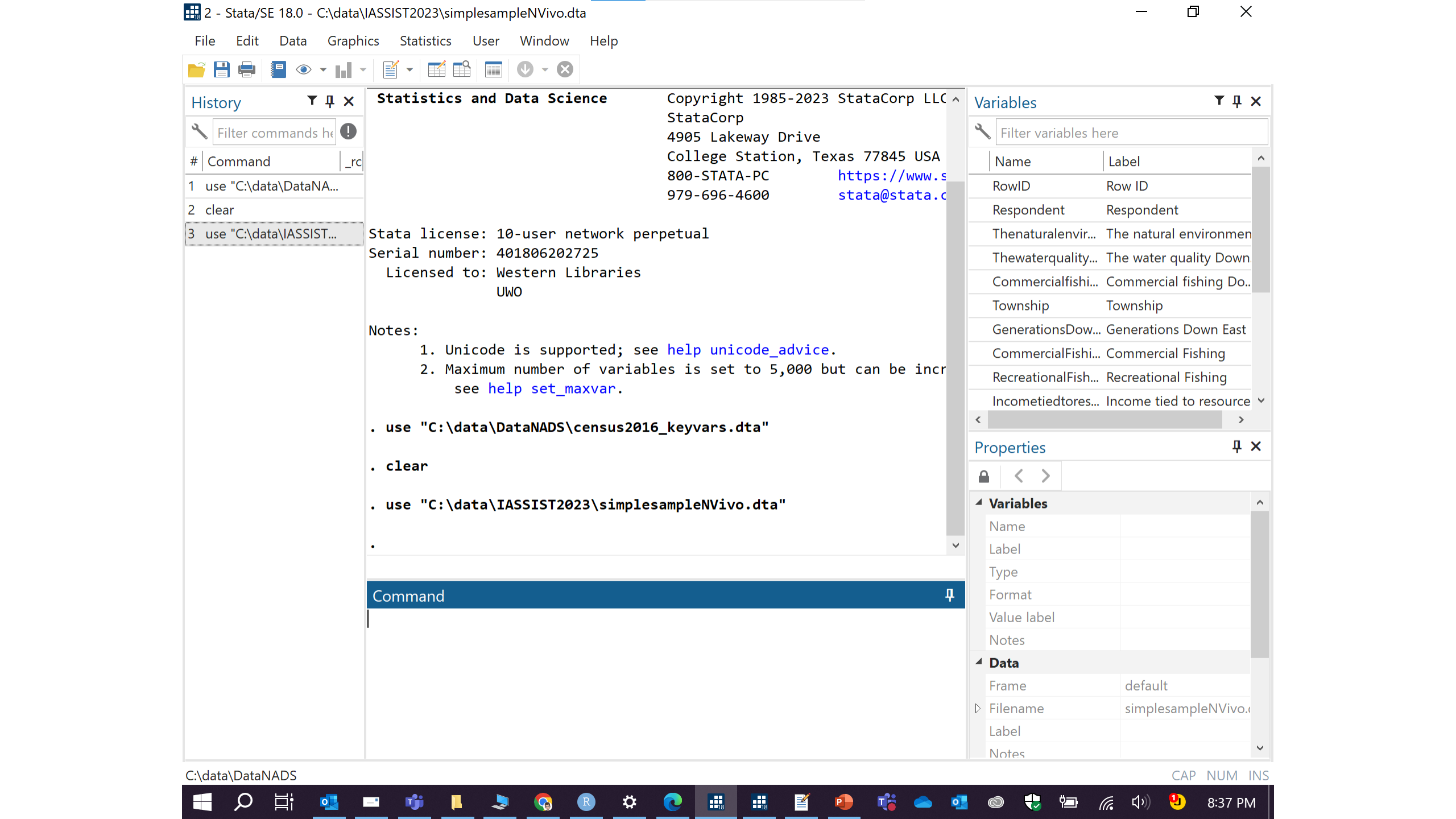
for a short break

Putting it into practice

A glimpse of Stata and R

Stata

Egen grouping, the equivalence class generator


2 - Stata/SE 18.0 - C:\data\IASSIST2023\simplesampleNVivo.dta
File Edit Data Graphics Statistics User Window Help
History
Filter commands he
Command _rc
1 use "C:\data\DataNA...
2 clear
3 use "C:\data\IASSIST...
Statistics and Data Science Copyright 1985-2023 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC https://www.s
979-696-4600 stata@stata.c
Stata license: 10-user network perpetual
Serial number: 401806202725
Licensed to: Western Libraries
UWO
Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5,000 but can be incr
see help set_maxvar.
. use "C:\data\DataNADS\census2016_keyvars.dta"
. clear
. use "C:\data\IASSIST2023\simplesampleNVivo.dta"
.
Command
Variables
Filter variables here
Name Label
RowID Row ID
Respondent Respondent
Thenaturalenvir... The natural environmen
Thewaterquality... The water quality Down
Commercialfishi... Commercial fishing Do..
Township Township
GenerationsDow... Generations Down East
CommercialFishi... Commercial Fishing
RecreationalFish... Recreational Fishing
Incometiedtores... Income tied to resource
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Frame default
Filename simplesampleNVivo.
Label
Notes
C:\data\DataNADS
CAP NUM INS
8:37 PM

3 - Stata/SE 18.0 - C:\data\IASSIST2023\simplesampleNVivo.dta

File Edit Data Graphics Statistics User Window Help

Some trade school/c..	6	9	15
Some undergraduate ..	3	8	11
Some undergraduate ..	0	1	1
Total	37	63	100

```
. tab EducationLevel Gender
```

Education Level	Gender Female	Male	Total
Completed graduate ..	1	3	4
Completed high school	5	12	17
Completed trade sch..	12	12	24
Completed undergrad..	8	11	19
Some graduate school	1	4	5
Some high school	1	3	4
Some trade school/c..	6	9	15
Some undergraduate ..	3	8	11
Some undergraduate ..	0	1	1
Total	37	63	100

```
.
```

Command

Variables

Filter variables here

Name	Label
RowID	Row ID
Respondent	Respondent
Thenaturalenvir...	The natural e
Thewaterquality...	The water qua
Commercialfishi...	Commercial fi
Township	Township
GenerationsDow...	Generations D
CommercialFishi...	Commercial F
RecreationalFish...	Recreational F
Incometiedtores...	Income tied t
Paceofdevelopm...	Pace of devel
Thetypesofdevel...	The types of
O	The types of
Age	Age
Gender	Gender
EducationLevel	Education Lev
Educ	group(Educati
education	

C:\data\DataNADS

CAP NUM INS

8:41 PM

3 - Stata/SE 18.0 - C:\data\IASSIST2023\simplesampleNVivo.dta

File Edit Data Graphics Statistics User Window Help

```
. egen testk_EducGender=group( Gender EducationLevel)
(4 missing values generated)

. tab testk_EducGender
```

group(Gender EducationLevel)	Freq.	Percent	Cum.
1	1	1.00	1.00
2	5	5.00	6.00
3	12	12.00	18.00
4	8	8.00	26.00
5	1	1.00	27.00
6	1	1.00	28.00
7	6	6.00	34.00
8	3	3.00	37.00
9	3	3.00	40.00
10	12	12.00	52.00
11	12	12.00	64.00
12	11	11.00	75.00
13	4	4.00	79.00
14	3	3.00	82.00
15	9	9.00	91.00
16	8	8.00	99.00
17	1	1.00	100.00
Total	100	100.00	

Variables

Filter variables here

Name	Label
RowID	Row ID
Respondent	Respondent
Thenaturalenvir...	The natural e
Thewaterquality...	The water qua
Commercialfishi...	Commercial fi
Township	Township
GenerationsDow...	Generations D
CommercialFishi...	Commercial F
RecreationalFish...	Recreational F
Incometiedtores...	Income tied t
Paceofdevelopm...	Pace of devel
Thetypesofdevel...	The types of
O	The types of
Age	Age
Gender	Gender
EducationLevel	Education Lev
Educ	group(Educati
education	
testk_EducGender	group(Gender

Command

C:\data\DataNADS

May 29, 2023
Monday

8:45 PM

3 - Stata/SE 18.0 - C:\data\IASSIST2023\simplesampleNVivo.dta

File Edit Data Graphics Statistics User Window Help

```
. tab education
```

education	Freq.	Percent	Cum.
High School or less	17	17.00	17.00
Some Post-Secondary	51	51.00	68.00
Completed University / College	32	32.00	100.00
Total	100	100.00	

```
.
```

Variables

Filter variables here

Name	Label
RowID	Row ID
Respondent	Respondent
Thenaturalenvir...	The natural e
Thewaterquality...	The water qua
Commercialfishi...	Commercial fi
Township	Township
GenerationsDow...	Generations D
CommercialFishi...	Commercial F
RecreationalFish...	Recreational F
Incometiedtores...	Income tied t
Paceofdevelopm...	Pace of devel
Thetypesofdevel...	The types of
O	The types of
Age	Age
Gender	Gender
EducationLevel	Education Lev
Educ	group(Educati
education	
testk_EducGender	group(Gender

Command

C:\data\DataNADS

CAP NUM INS

9:08 PM

3 - Stata/SE 18.0 - C:\data\IASSIST2023\simplesampleNVivo.dta

File Edit Data Graphics Statistics User Window Help

Gender	education High Scho	Some Post	Completed	Total
Female	5	21	11	37
Male	12	30	21	63
Total	17	51	32	100

```
. egen testk_EducGender=group( Gender education )
(4 missing values generated)

. tab testk_EducGender
```

group(Gender education)	Freq.	Percent	Cum.
1	5	5.00	5.00
2	21	21.00	26.00
3	11	11.00	37.00
4	12	12.00	49.00
5	30	30.00	79.00
6	21	21.00	100.00
Total	100	100.00	

```
.
```

Variables

Filter variables here

Name	Label
RowID	Row ID
Respondent	Respondent
Thenaturalenvir...	The natural e
Thewaterquality...	The water qua
Commercialfishi...	Commercial fi
Township	Township
GenerationsDow...	Generations D
CommercialFishi...	Commercial F
RecreationalFish...	Recreational F
Incometiedtores...	Income tied t
Paceofdevelopm...	Pace of devel
Thetypesofdevel...	The types of
O	The types of
Age	Age
Gender	Gender
EducationLevel	Education Lev
Educ	group(Educati
education	
testk_EducGender	group(Gender

Command

C:\data\DataNADS

CAP NUM INS

9:10 PM

3 - Stata/SE 18.0 - C:\data\IASSIST2023\simplesampleNVivo.dta

File Edit Data Graphics Statistics User Window Help

```
. egen testk_4=group( Gender education Age Township )
(4 missing values generated)

. tab testk_4
```

group(Gender education Age Township)	Freq.	Percent	Cum.
1	1	1.00	1.00
2	1	1.00	2.00
3	1	1.00	3.00
4	1	1.00	4.00
5	1	1.00	5.00
6	1	1.00	6.00
7	1	1.00	7.00
8	1	1.00	8.00
9	1	1.00	9.00
10	1	1.00	10.00
11	1	1.00	11.00
12	1	1.00	12.00
13	1	1.00	13.00
14	1	1.00	14.00
15	1	1.00	15.00
16	1	1.00	16.00

Variables

Filter variables here

Name	Label
RowID	Row ID
Respondent	Respondent
Thenaturalenvir...	The natural e
Thewaterquality...	The water qua
Commercialfishi...	Commercial fi
Township	Township
GenerationsDow...	Generations D
CommercialFishi...	Commercial F
RecreationalFish...	Recreational F
Incometiedtores...	Income tied t
Paceofdevelopm...	Pace of devel
Thetypesofdevel...	The types of
O	The types of
Age	Age
Gender	Gender
EducationLevel	Education Lev
Educ	group(Educati
education	
testk_EducGender	group(Gender
testk_4	group(Gender

Command

C:\data\DataNADS

CAP NUM INS

9:12 PM

R, RStudio

The SDCMicro package

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

indRisk_table | SDCWorkshopDemoR.R*

Filter

	district	hhh_ethnolinguistic_group	hhh_marital_status	hhh_age	hhh_gender	hhh_disability	tota
1	Khuram Sarbagh	Arab	Married	47	Male	no	
2	Sar e Pul	Arab	Married	54	Female	no	
3	Pul e khumri	Arab	Married	30	Male	no	
4	Dawlatabad	Arab	Married	33	Female	no	
5	Sholgareh	Baloch	Divorced	33	Male	no	
6	Khulm	Arab	Married	36	Male	no	
7	Sholgareh	Arab	Married	56	Male	no	
8	Khuram Sarbagh	Arab	Single	46	Male	no	
9	Khuram Sarbagh	Arab	Married	48	Male	no	
10	Khulm	Arab	Married	55	Male	yes	

Showing 1 to 10 of 7,005 entries, 11 total columns

Console | Terminal | Background Jobs

Terminal 1

```
Microsoft Windows [Version 10.0.19045.2965]
(c) Microsoft Corporation. All rights reserved.

c:\data\IASSIST2023>
```

Environment

```
indRisk_table...
View(indRisk_...
print(objSDC,...
objSDC <- glo...
table(objSDC@...
print(objSDC,...
print(objSDC,...
```

Files | Plots

Zoom

8:49 PM

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

indRisk_table | SDCWorkshopDemoR.R

Filter

	district	hhh_ethnolinguistic_group	hhh_marital_status	hhh_age	hhh_gender	hhh_disability	total_members
1	Khuram Sarbagh	Arab	Married	47	Male	no	
2	Sar e Pul	Arab	Married	54	Female	no	
3	Pul e khumri	Arab	Married	30	Male	no	
4	Dawlatabad	Arab	Married	33	Female	no	
5	Sholgareh	Baloch	Divorced	33	Male	no	
6	Khulm	Arab	Married	36	Male	no	
7	Sholgareh	Arab	Married	56	Male	no	
8	Khuram Sarbagh	Arab	Single	46	Male	no	
9	Khuram Sarbagh	Arab	Married	48	Male	no	

Showing 1 to 9 of 7,005 entries, 11 total columns

Environment

install.package...

Files Plots

Zoom

Console Background Jobs

R 4.3.0 · ~/

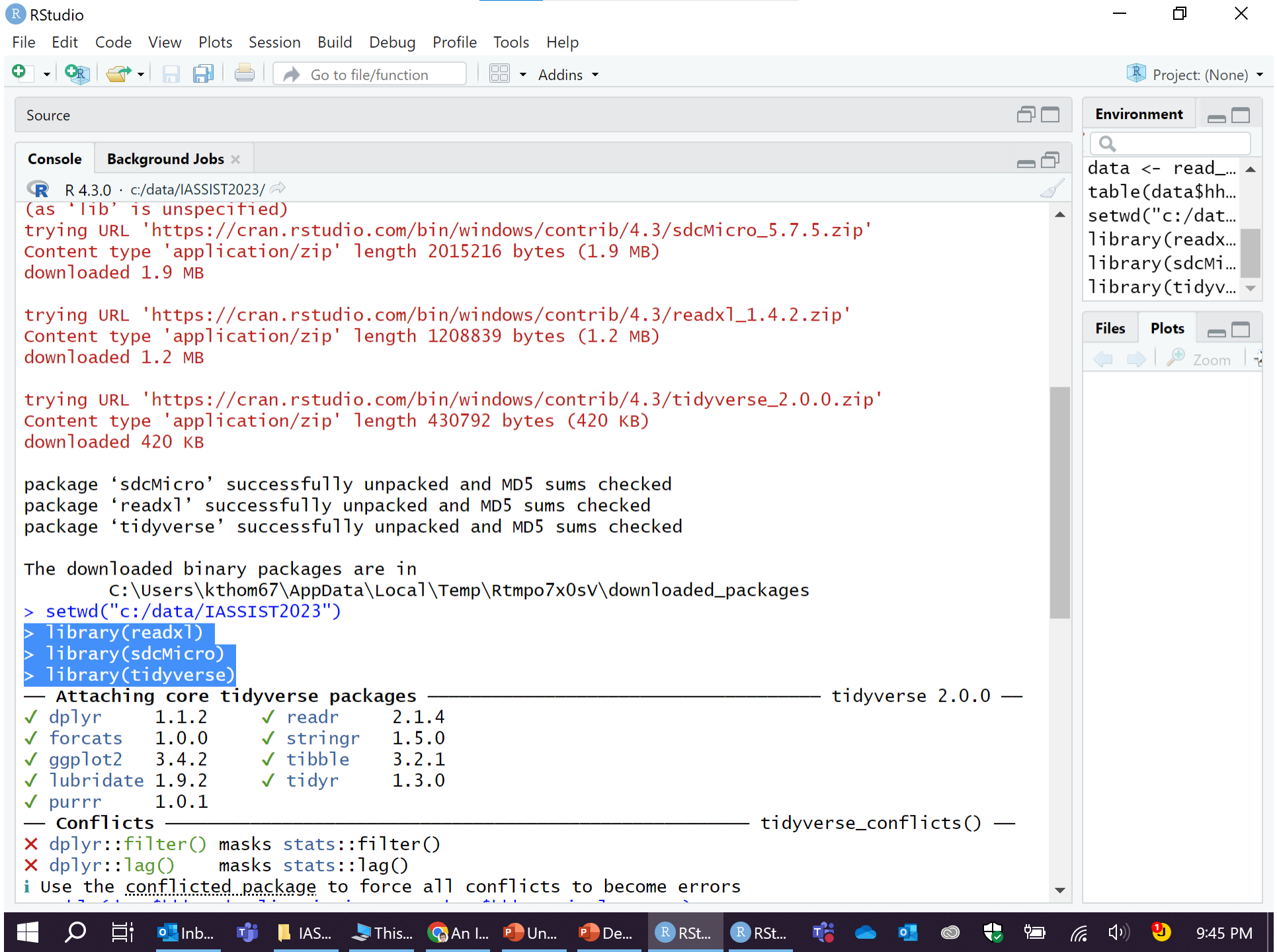
```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages(c("sdcMicro", "readxl", "tidyverse"))
```

Inbox... | IASSI... | This ... | An In... | Unde... | Dem... | RStu... | 9:35 PM


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Background Jobs
R 4.3.0 · c:/data/IASSIST2023/
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/sdcMicro_5.7.5.zip'
Content type 'application/zip' length 2015216 bytes (1.9 MB)
downloaded 1.9 MB
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/readxl_1.4.2.zip'
Content type 'application/zip' length 1208839 bytes (1.2 MB)
downloaded 1.2 MB
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/tidyverse_2.0.0.zip'
Content type 'application/zip' length 430792 bytes (420 KB)
downloaded 420 KB
package 'sdcMicro' successfully unpacked and MD5 sums checked
package 'readxl' successfully unpacked and MD5 sums checked
package 'tidyverse' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
C:\Users\kthom67\AppData\Local\Temp\Rtmpo7x0sV\downloaded_packages
> setwd("c:/data/IASSIST2023")
> library(readxl)
> library(sdcMicro)
> library(tidyverse)
── Attaching core tidyverse packages ──── tidyverse 2.0.0 ──
✓ dplyr 1.1.2 ✓ readr 2.1.4
✓ forcats 1.0.0 ✓ stringr 1.5.0
✓ ggplot2 3.4.2 ✓ tibble 3.2.1
✓ lubridate 1.9.2 ✓ tidyr 1.3.0
✓ purrr 1.0.1
── Conflicts ──── tidyverse_conflicts() ──
✕ dplyr::filter() masks stats::filter()
✕ dplyr::lag() masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
Environment
data <- read_...
table(data$hh...
setwd("c:/dat...
library(readx...
library(sdcMi...
library(tidyv...
Files
Plots
Zoom
9:45 PM

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Source

Console | Background Jobs

R 4.3.0 · c:/data/IASSIST2023/

```
> table(data$hhh_ethnolinguistic_group,data$hhh_marital_status)

           Divorced Married Single Widowed
  Arab          192    6132    206      15
  Aymaq           1      44      0       0
  Baloch          3      39      3       0
  Brahui          0       5      0       0
  Gujjar          1      42      2       0
  Hazara          1      42      2       0
  Kyrgyz          0       5      0       0
  Nuristani       1      44      0       0
  Pamiri          1       4      0       0
  Pashai          0      43      2       0
  Pashtun         1      39      0       0
  Tajik           1      44      0       0
  Turkmen         1      42      2       0
  Uzbek           2      40      3       0
>
```

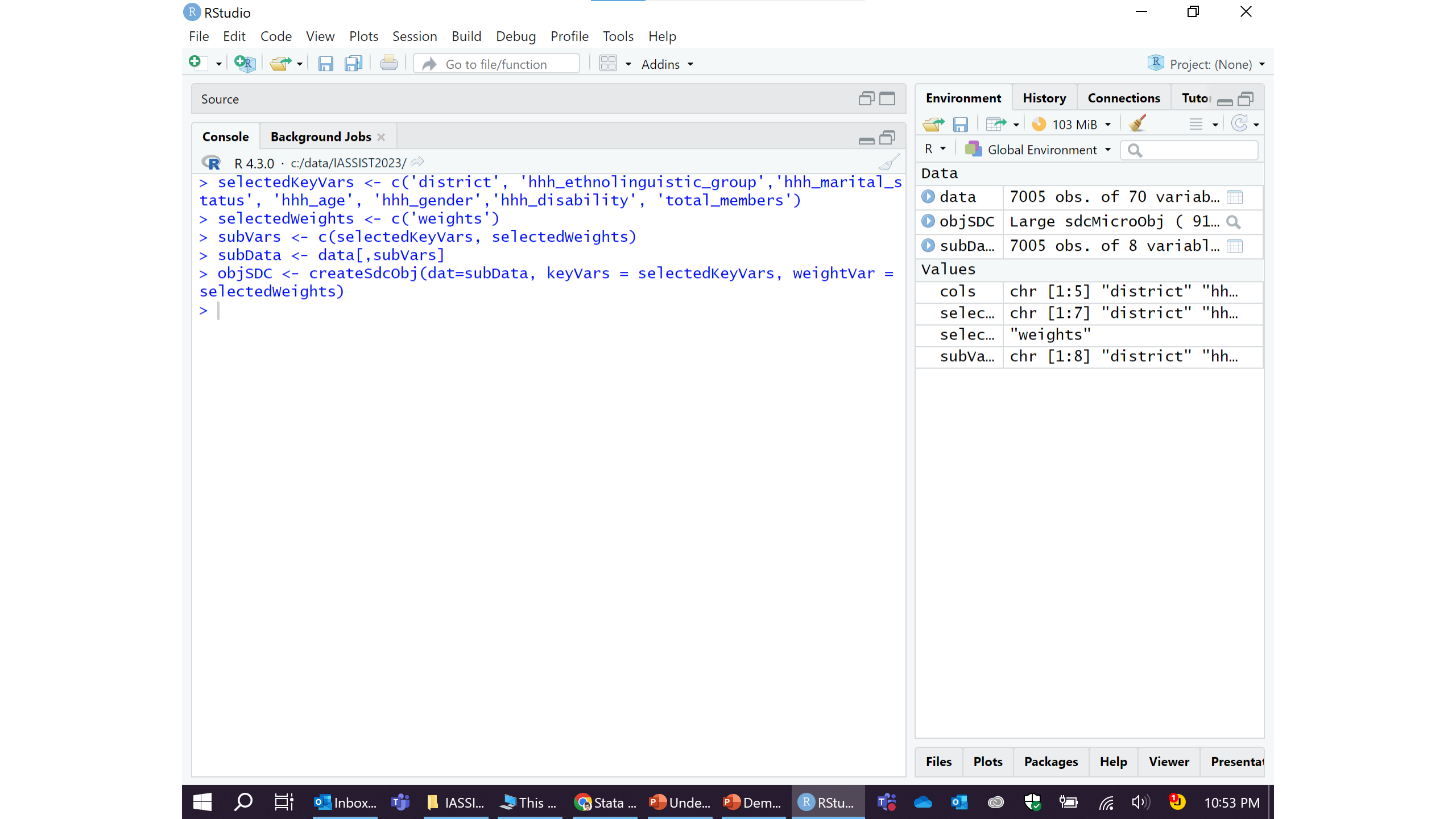
Environment

```
setwd("c:/dat...
library(readx...
library(sdcMi...
library(tidyv...
data <- read_...
table(data$hh...
```

Files | Plots

Zoom

Inb... IAS... This... clea... Un... De... RSt... RSt... 9:47 PM


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Source
Console
Background Jobs
R 4.3.0 · c:/data/IASSIST2023/
> selectedKeyVars <- c('district', 'hhh_ethnolinguistic_group','hhh_marital_status', 'hhh_age', 'hhh_gender','hhh_disability', 'total_members')
> selectedWeights <- c('weights')
> subVars <- c(selectedKeyVars, selectedWeights)
> subData <- data[,subVars]
> objSDC <- createSdcObj(dat=subData, keyVars = selectedKeyVars, weightVar = selectedWeights)
>
Environment
History
Connections
Tuto
103 MiB
R
Global Environment
Data
data 7005 obs. of 70 variab...
objSDC Large sdcMicroObj (91...
subDa... 7005 obs. of 8 variabl...
Values
cols chr [1:5] "district" "hh...
selec... chr [1:7] "district" "hh...
selec... "weights"
subVa... chr [1:8] "district" "hh...
Files
Plots
Packages
Help
Viewer
Presenta
Inbox...
IASSI...
This ...
Stata ...
Unde...
Dem...
RStu...
10:53 PM

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Source

Console | Background Jobs

R 4.3.0 · c:/data/IASSIST2023/

```
> print(objSDC, type="kAnon")
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 3255 (46.467%)
  - 3-anonymity: 4599 (65.653%)
  - 5-anonymity: 5873 (83.840%)

-------------------------------------------------------------------------------


> 
```

Environment | History | Connections | Tuto

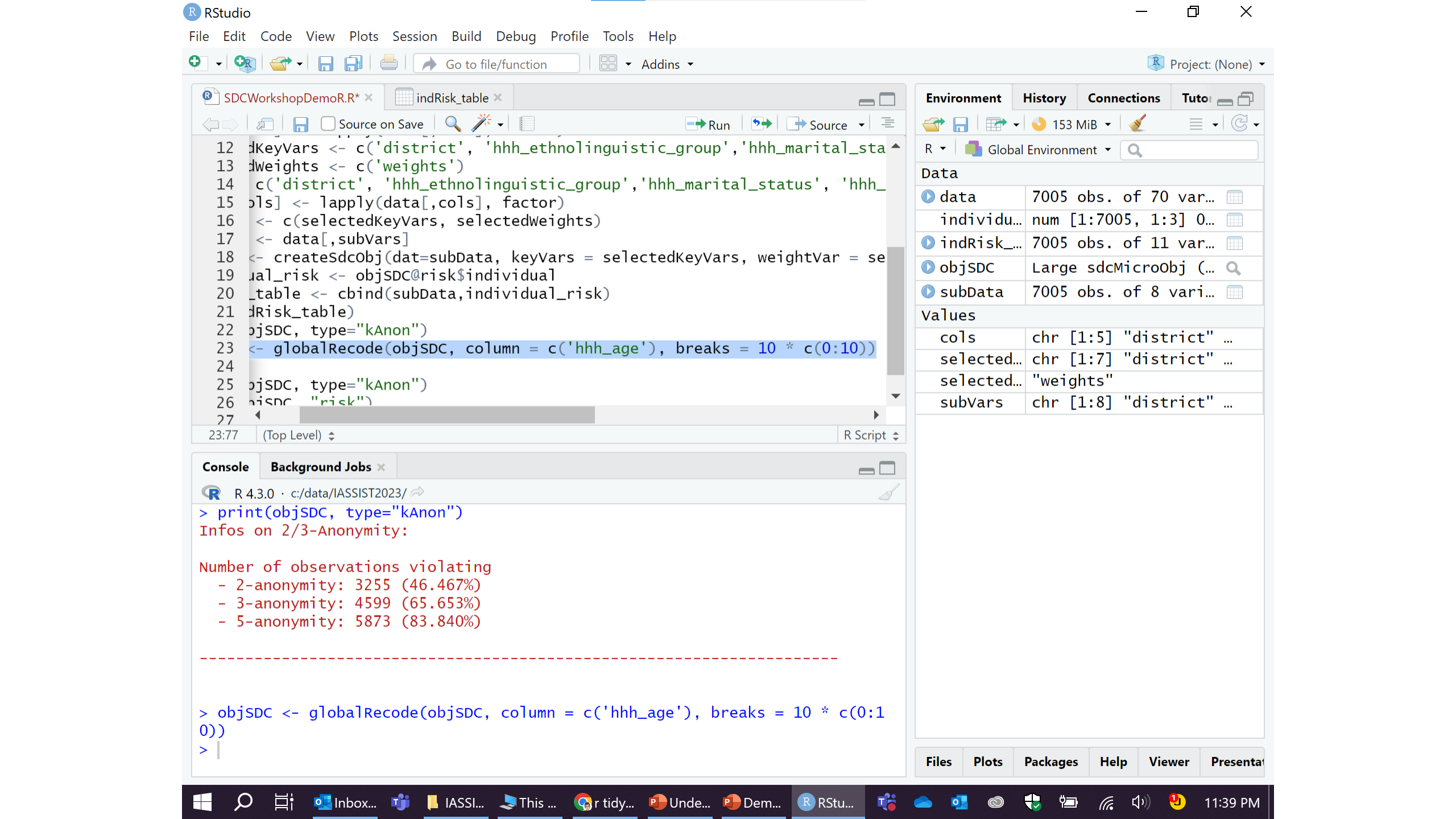
146 MiB

R | Global Environment

Data	
data	7005 obs. of 70 var…
individu…	num [1:7005, 1:3] 0…
indRisk_…	7005 obs. of 11 var…
objSDC	Large sdcMicroObj (…
subData	7005 obs. of 8 vari…
Values	
cols	chr [1:5] "district" …
selected…	chr [1:7] "district" …
selected…	"weights"
subVars	chr [1:8] "district" …

Files | Plots | Packages | Help | Viewer | Presenta

Inbox… | IASSI… | This … | r tidy… | Unde… | Dem… | RStu… | 11:30 PM


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
SDCWorkshopDemoR.R*
indRisk_table
Source on Save
Run
Source
dKeyVars <- c('district', 'hhh_ethnolinguistic_group','hhh_marital_sta
dWeights <- c('weights')
c('district', 'hhh_ethnolinguistic_group','hhh_marital_status', 'hhh_
ols] <- lapply(data[,cols], factor)
<- c(selectedKeyVars, selectedWeights)
<- data[,subVars]
<- createSdcObj(dat=subData, keyVars = selectedKeyVars, weightVar = se
ual_risk <- objSDC@risk$individual
_table <- cbind(subData,individual_risk)
dRisk_table)
ojSDC, type="kAnon")
<- globalRecode(objSDC, column = c('hhh_age'), breaks = 10 * c(0:10))
ojSDC, type="kAnon")
23:77
(Top Level)
R Script
Console
Background Jobs
R 4.3.0 · c:/data/IASSIST2023/
> print(objSDC, type="kAnon")
Infos on 2/3-Anonymity:
Number of observations violating
- 2-anonymity: 3255 (46.467%)
- 3-anonymity: 4599 (65.653%)
- 5-anonymity: 5873 (83.840%)
> objSDC <- globalRecode(objSDC, column = c('hhh_age'), breaks = 10 * c(0:1
0))
>
Environment
History
Connections
153 MiB
R
Global Environment
Data
data 7005 obs. of 70 var…
individu… num [1:7005, 1:3] 0…
indRisk_… 7005 obs. of 11 var…
objSDC Large sdcMicroObj (…
subData 7005 obs. of 8 vari…
Values
cols chr [1:5] "district" …
selected… chr [1:7] "district" …
selected… "weights"
subVars chr [1:8] "district" …
Files
Plots
Packages
Help
Viewer
Inbox…
IASSI…
This …
r tidy…
Unde…
Dem…
RStu…
11:39 PM

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Source

Console | Background Jobs

R 4.3.0 · c:/data/IASSIST2023/

```
> print(objSDC, type="kAnon")
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 3255 (46.467%)
  - 3-anonymity: 4599 (65.653%)
  - 5-anonymity: 5873 (83.840%)

---------------------------------------------------------------------


> objSDC <- globalRecode(objSDC, column = c('hhh_age'), breaks = 10 * c(0:1
0))
> print(objSDC, type="kAnon")
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 1345 (19.201%) | in original data: 3255 (46.467%)
  - 3-anonymity: 1955 (27.909%) | in original data: 4599 (65.653%)
  - 5-anonymity: 2710 (38.687%) | in original data: 5873 (83.840%)

---------------------------------------------------------------------


>
```

Environment | History | Connections | Tuto

155 MiB

R | Global Environment

Data	
data	7005 obs. of 70 var…
individu…	num [1:7005, 1:3] 0…
indRisk_…	7005 obs. of 11 var…
objSDC	Large sdcMicroObj (…
subData	7005 obs. of 8 vari…
Values	
cols	chr [1:5] "district" …
selected…	chr [1:7] "district" …
selected…	"weights"
subVars	chr [1:8] "district" …

Files | Plots | Packages | Help | Viewer | Presenta

Inbox… | IASSI… | This … | r tidy… | Unde… | Dem… | RStu… | 11:39 PM

The end… so far

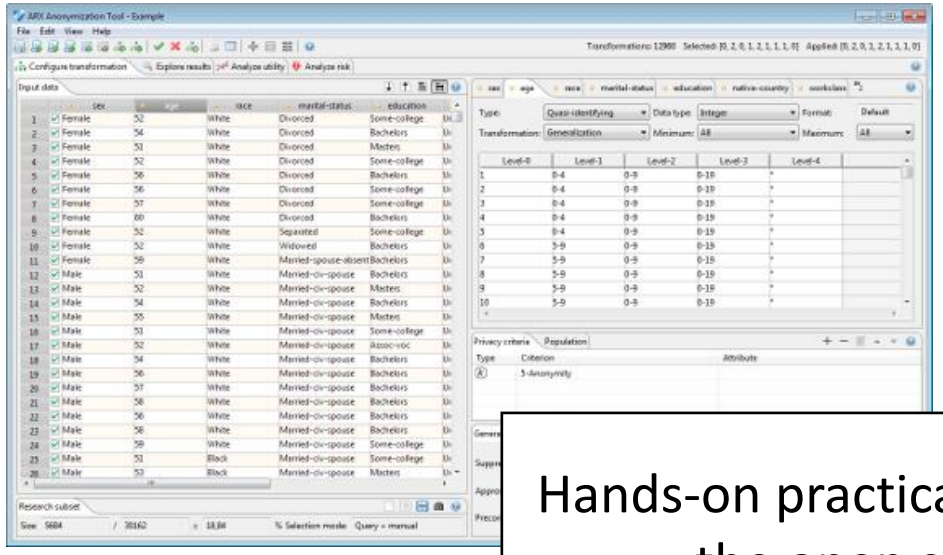
- Questions, comments, concerns, confusions?

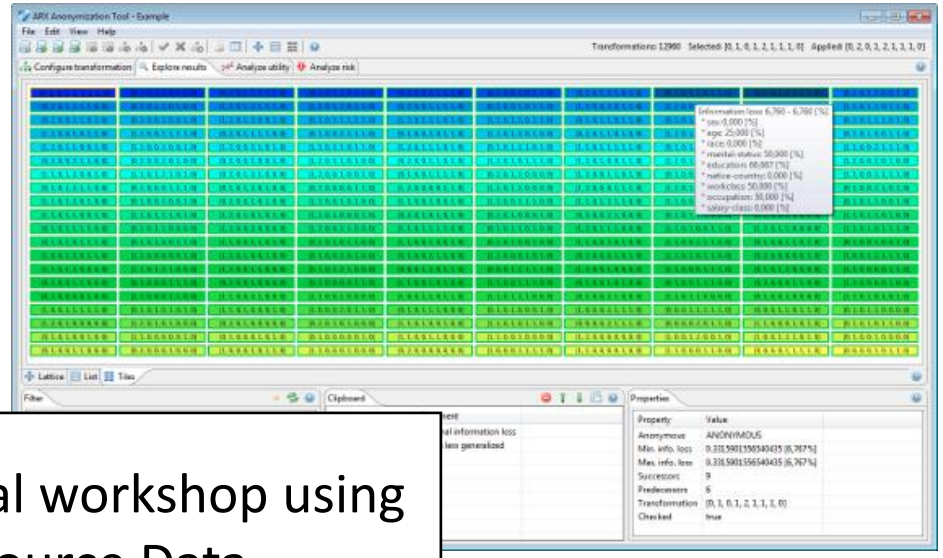
- Kristi Thompson, *kthom67@uwo.ca*

- AI images created with NightCafe Creator, using the DALL·E 2 and Stable Diffusion algorithms

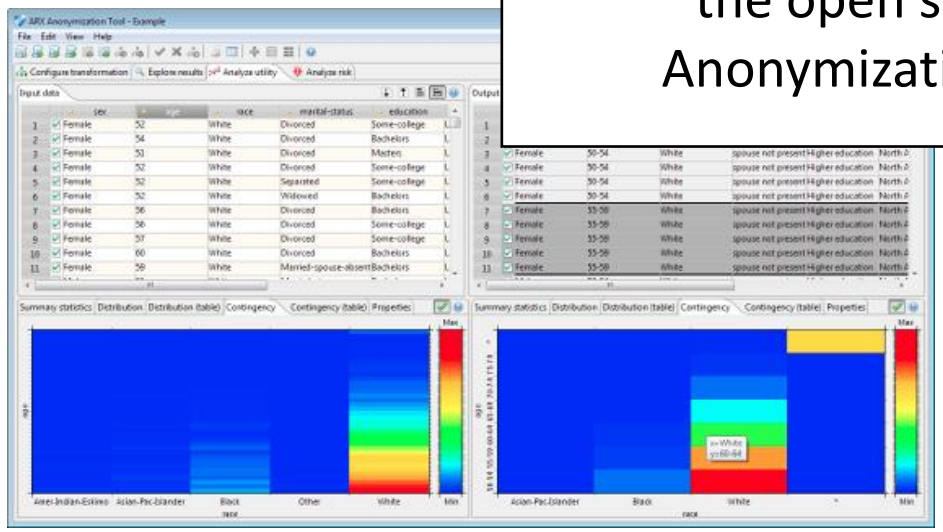
Western UNIVERSITY · CANADA

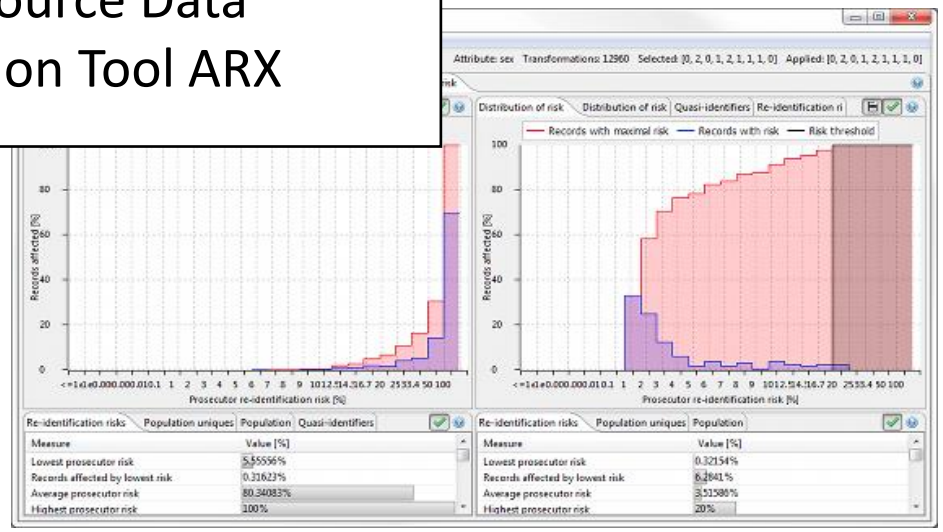
Next Workshop:





Hands-on practical workshop using the open source Data Anonymization Tool ARX





Wednesday October 4, 2023, 10:00am

Sources and further reading

Ayala-Rivera, V., McDonagh, P., Cerqueus, T. and Murphy, L. (2014) 'A systematic comparison and evaluation of k-anonymization algorithms for practitioners', Transactions on data privacy, 7(3), pp.337-370.

British Medical Journal. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ* 2010; 340. Available at https://doi.org/10.1136/bmj.c181

Domingo-Ferrer, J. and Torra, V. (2008) 'A critique of k-anonymity and some of its enhancements', In Third International Conference on Availability, Reliability and Security. IEEE.

Elliot, M., Mackey, E., O'Hara, K. and Tudor, C. (2016) *The Anonymisation Decision-Making Framework*, Manchester: UKAN. Available at https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

Humdata.org. Introduction to Disclosure Risk Assessment and SDCMicro

Portage Covid-19 Working Group. 'De-identification Guidance', available at https://zenodo.org/record/4270551#.Ygvklt_MKM8

Samarati, P. and Sweeney, L. (1998) 'Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression', available at https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf

Thompson, K and Sullivan, C. (2020) 'Mathematics, risk and messy survey data', IASSIST Quarterly 44 (4), available at https://iassistquarterly.com/index.php/iassist/article/download/979/961

Rod, Alisa and Kristi Thompson, 2023. Sensitive Data: Practical and Theoretical Considerations. In Research Data Management in the Canadian Context: a Guide for practitioners and Learners.