pathways

Pathways Segmentation Methods Workshop

Day 1 – Overview and Data Cleaning

IDM Gates Foundation

Day 1 Outline

- Pathways approach
- Data segmentation
- Vocabulary
- Cluster analysis
- Pathways segmentation process
- Identifying vulnerability factors

Day 1 Outline

- Step 3 data cleaning
- Technical guide
- Activity: Pathways Workbook in Excel & R code
- Activity: Setting up the R environment
- Activity: Data cleaning and univariate code

Pathways

Understanding social, cultural, economic, and environmental vulnerability to improve women's health & wellbeing

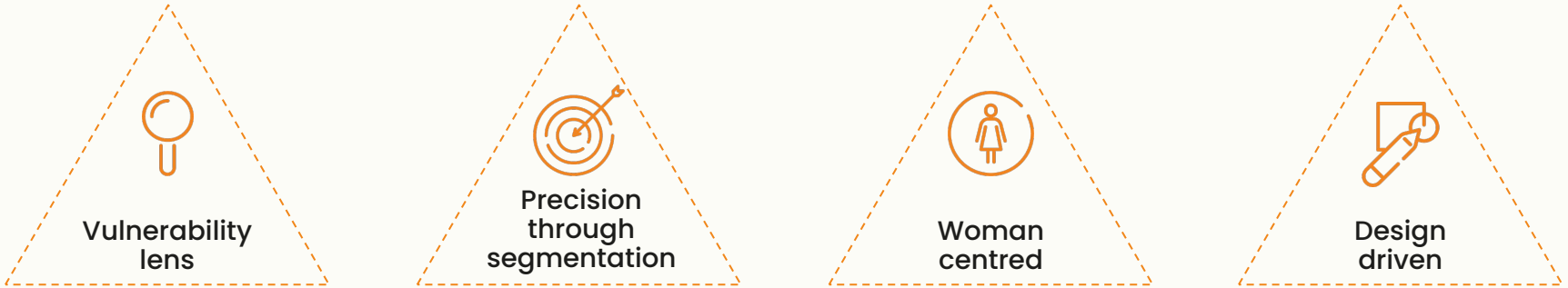
www.projectpathways.org

01

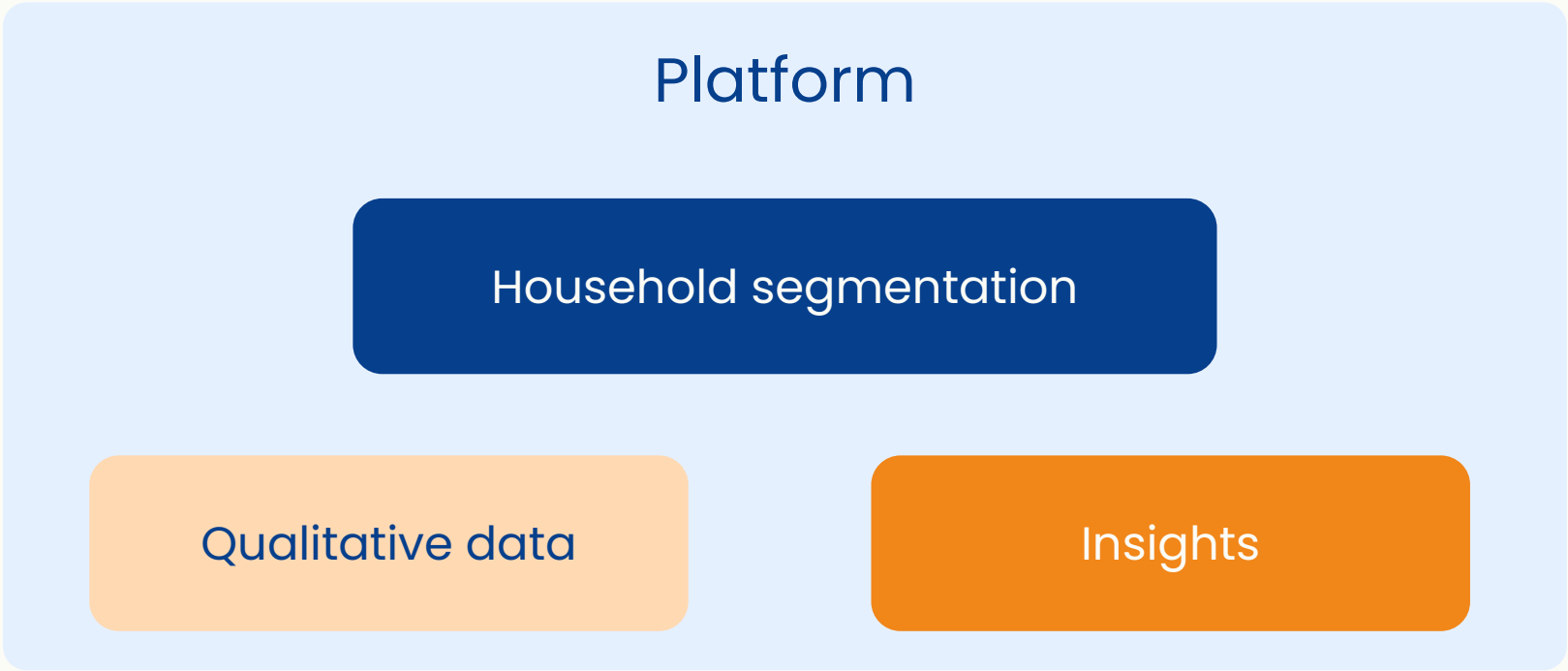
Pathways Approach

Pathways provides data, insights and tools that make the social determinants of health **actionable** and **accelerate better health**.

FROM			
CLINICAL CARE ONLY interventions that neglect the root causes of illness	**ONE SIZE FITS ALL** approaches that overlook diverse needs	**TOP DOWN VERTICAL** programs dictated mostly by guidelines	**DISCONNECTION** between health systems and other social systems
TO			
SOCIAL MEDICINE that addresses upstream social conditions and inequities driving health outcomes.	**DIFFERENTIATED SERVICES** that identify and prioritize the needs of vulnerable populations	**HORIZONTAL, HUMAN CENTRED** services that holistically promote well-being and adapt to human needs	**INTEGRATION** of health programs with social and community systems to enhance impact



Pathways Approach



How does it all **come together?**

Vulnerability framework

- Conceptual Framework
- Factors & relationships
- Analytic approach

Data & Insights

Quantitative + Qualitative

- Segmentations
- Insights and Journeys
- Data sets for further analysis

Methods & tools

- Technical assistance
- Typing tool
- Research & Design tools

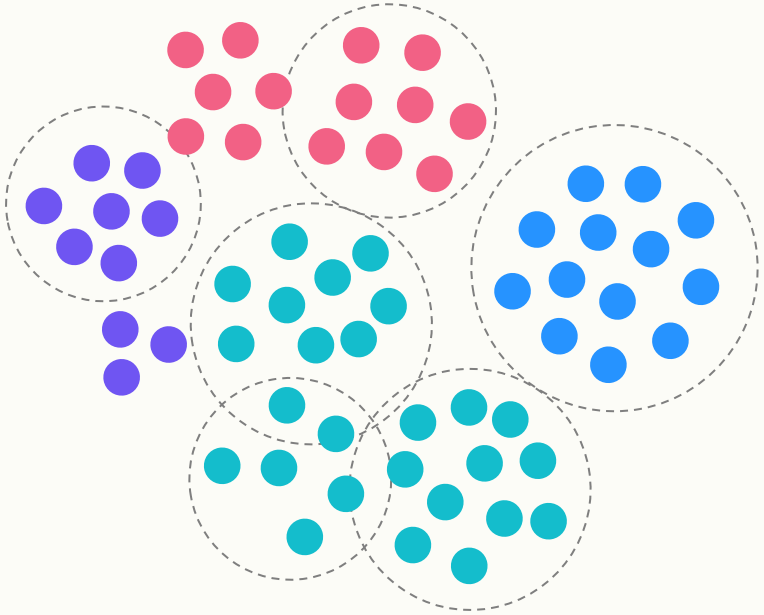
Platform

How we define **Vulnerability**

The quality or state of potentially being harmed, either physically, socially, cognitively, or emotionally, related to reproductive, newborn, maternal and child health and nutrition.

Vulnerability and resilience are not opposite ends of one spectrum.

Pathways Approach



People accessing the health system

- Most vulnerable
- More vulnerable
- Less vulnerable
- Least vulnerable

Segments cluster households according to their vulnerability using key **social, cultural, economic, and environmental indicators** closely associated with RMNCH outcomes.

Pathways Approach

Pathways provides a holistic view of women and their households' health and wellbeing to operationalize the social determinants of health

Pathways Approach

Holistic concept of health and wellbeing

Social determinants of health

Vulnerability as a tool to address risk before harm

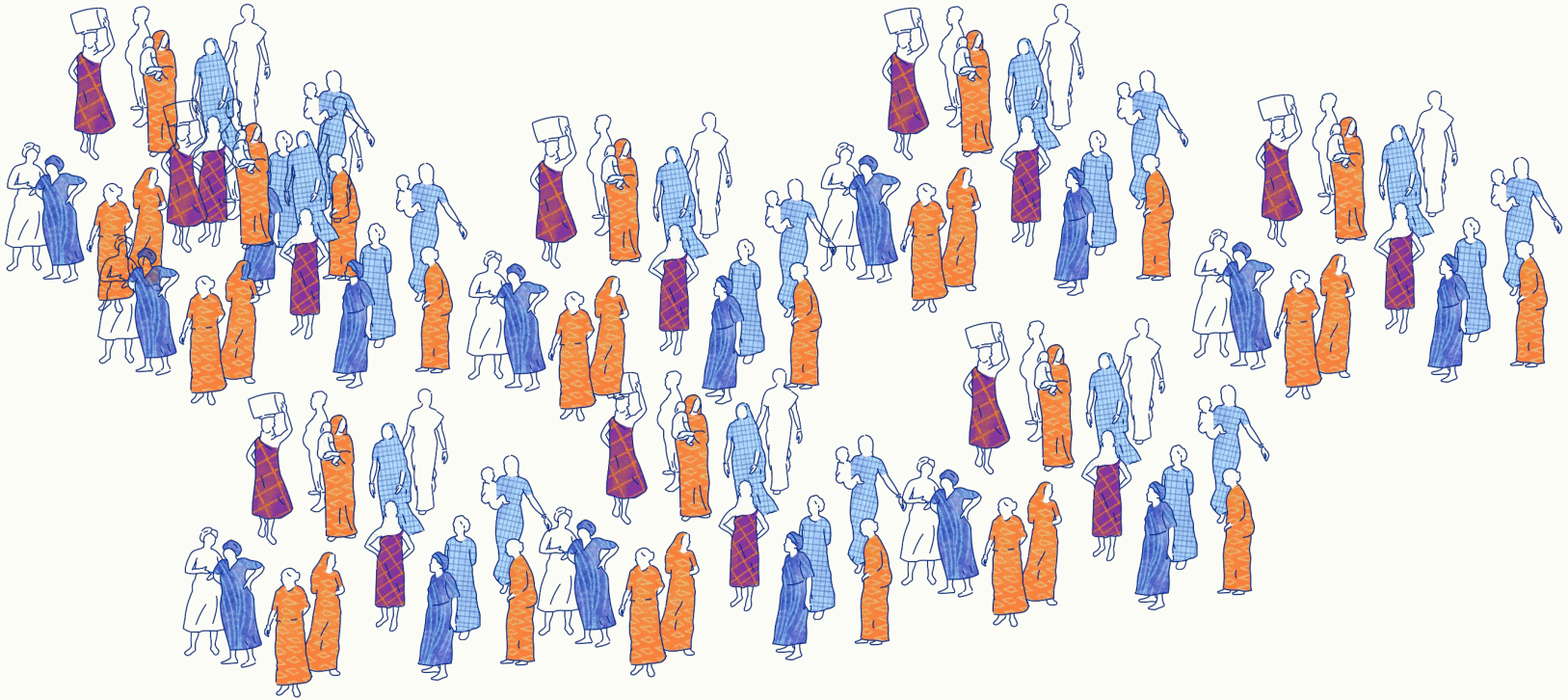
02

Data Segmentation

segmentation:

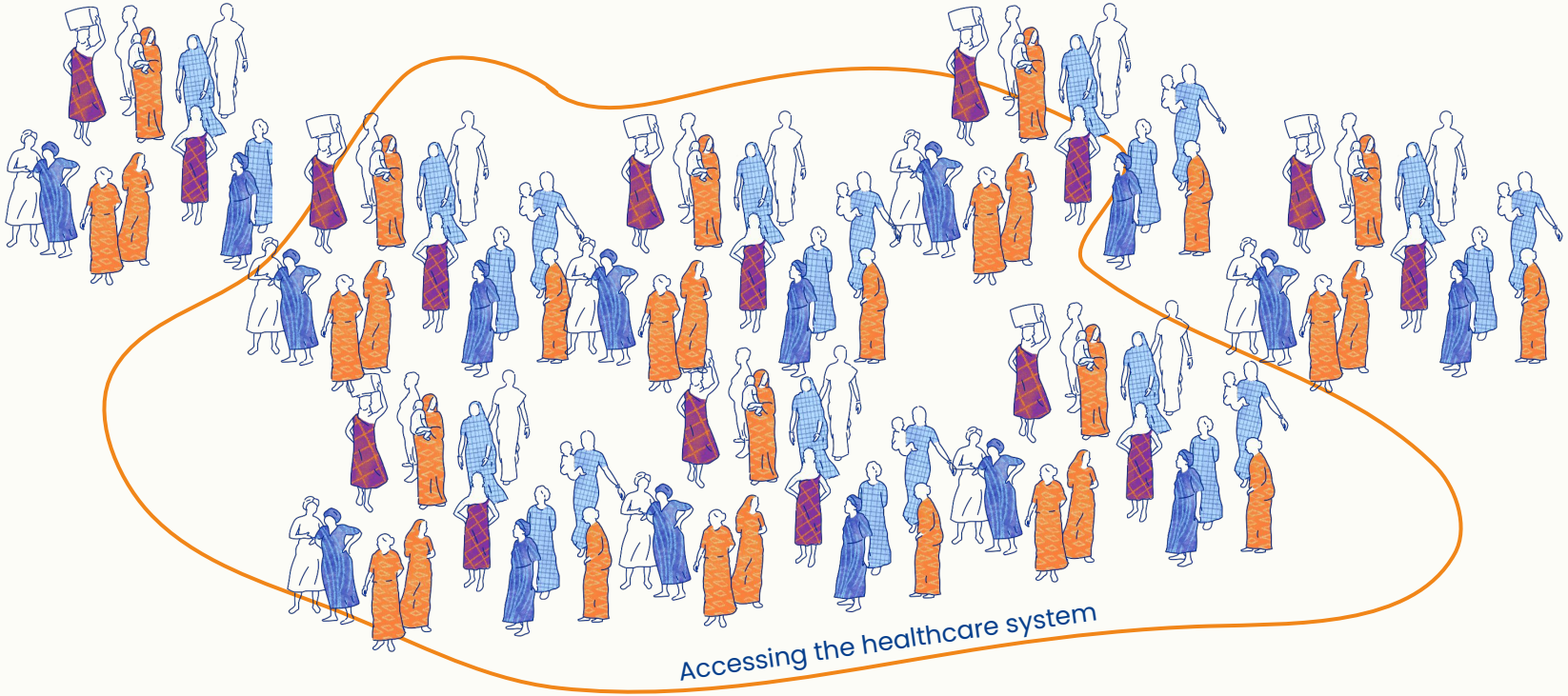
the process of dividing a population into smaller sub-groups (known as segments) based on shared characteristics

Imagine 120 women and their stories

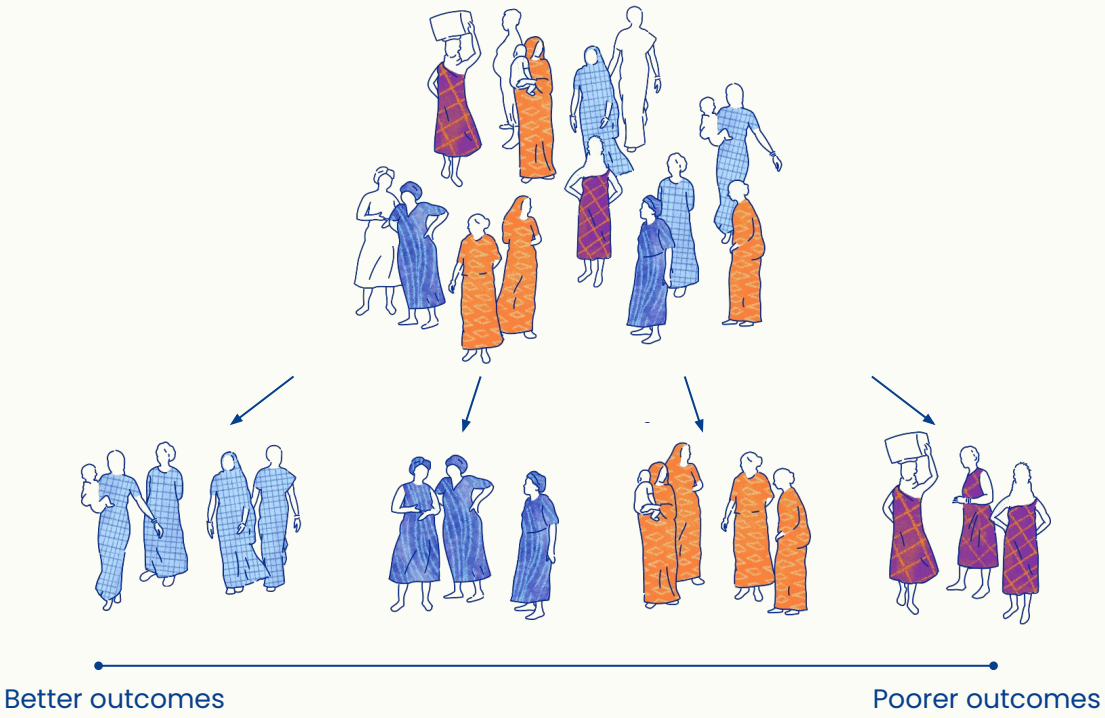


The **traditional** global health approach

Focuses predominantly on reaching a population with **one size fits all** approaches within the health system



What is a **segmentation?**



Dividing a population into smaller sub-groups known as segments.

Sub-groups can be based on **demographics, behaviors, geography, needs, and other factors.**

Segments are mutually exclusive, i.e. a person can only belong to one segment.

Segments form an order, where segments rank according to a specific dimension of interest.

How do they differ among a population?

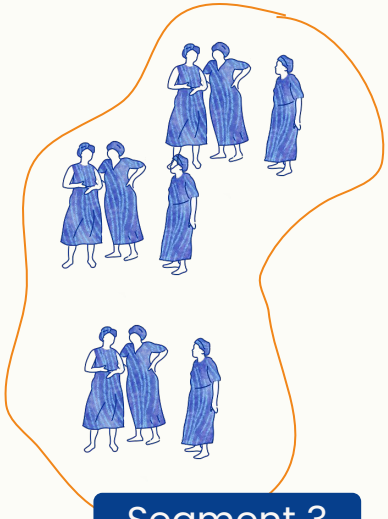
Accessing the healthcare system

Segment 1

Poorest
health outcomes

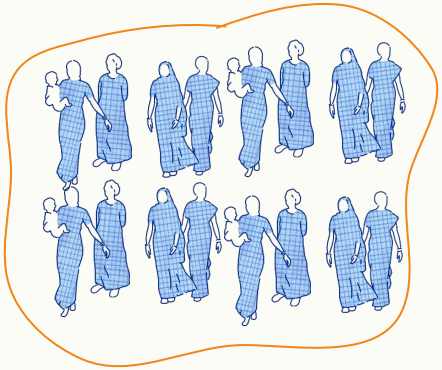
Segment 2

Poor
health outcomes



Segment 3

Better
health outcomes



Segment 4

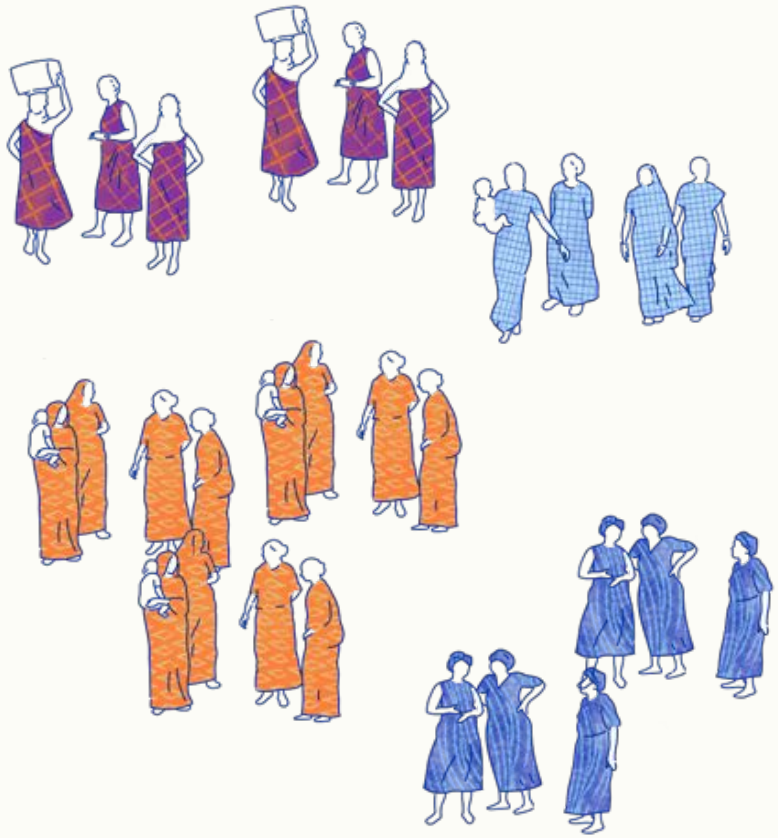
Best
health outcomes

Demographics

Geography

Needs

Behavioral factors



Identify homogeneous “targets” for the creation/ provision of services, products, interventions, communication campaigns, etc.

03

Vocabulary

domain:

category of related factors that describe similar aspects of women's vulnerability

The woman & her past experience

Household relationships

Household economics & living conditions

Social support

Health mental models

Human & natural systems

factor:

a specific fact, situation, or construct that help define how women experience vulnerability

measures (or variables):

a standardized way of quantifying factors so they can be used in statistical analysis

segment:

a population
sub-group created
using cluster analysis

cluster analysis:

a statistical method used to organize observations into meaningful groups – or clusters – that share common characteristics

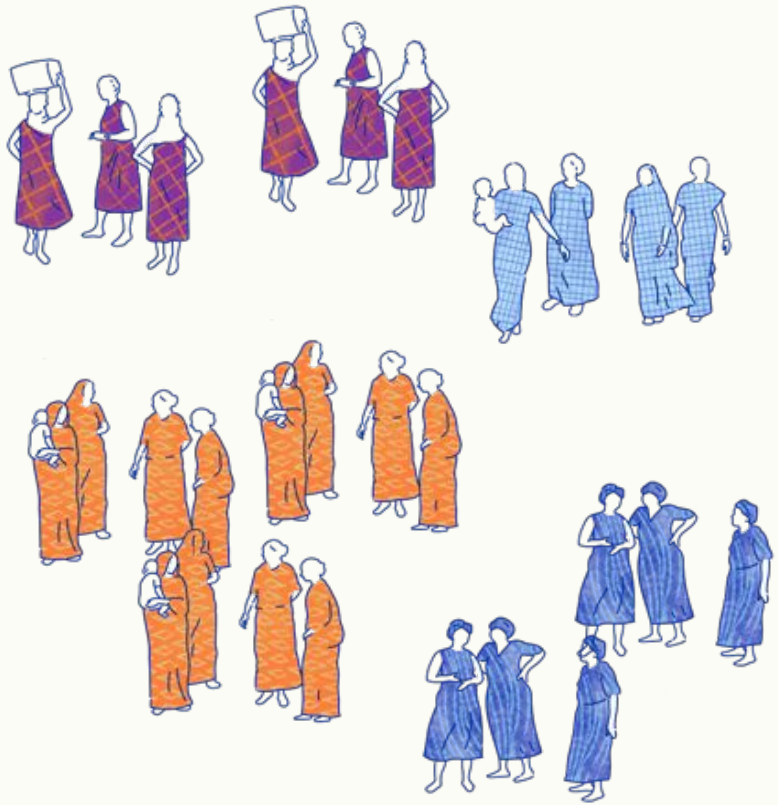
segmentation solution:

the final set of segments resulting from completing the Pathways segmentation methodology

04

Cluster Analysis

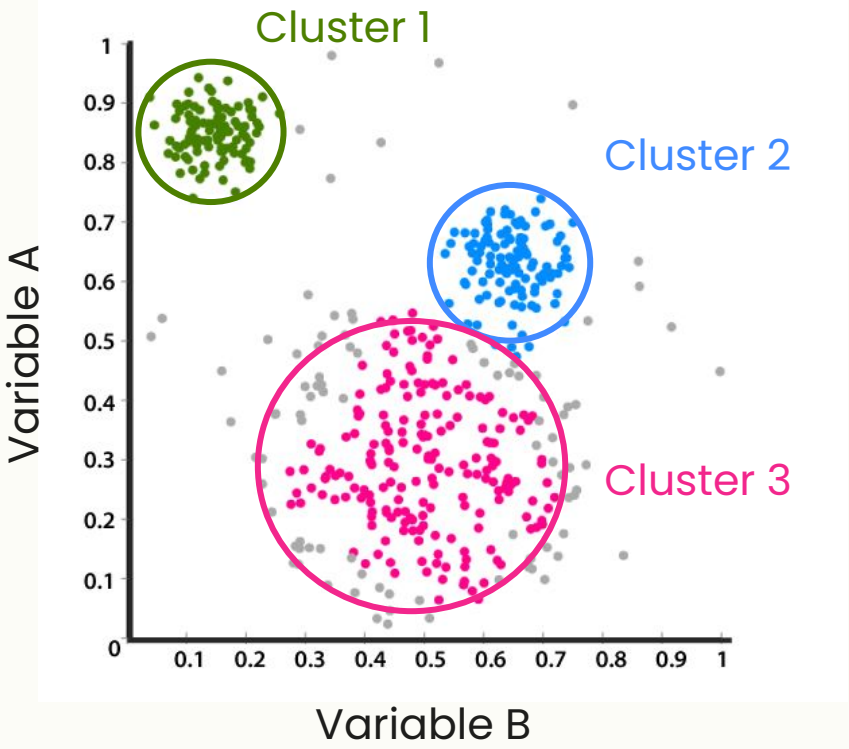
The objective of cluster analysis is to find similar groups of subjects



Observations within each group are similar to one another with respect to variables or attributes of interest

Observations between the groups are very different from one another

A simple example

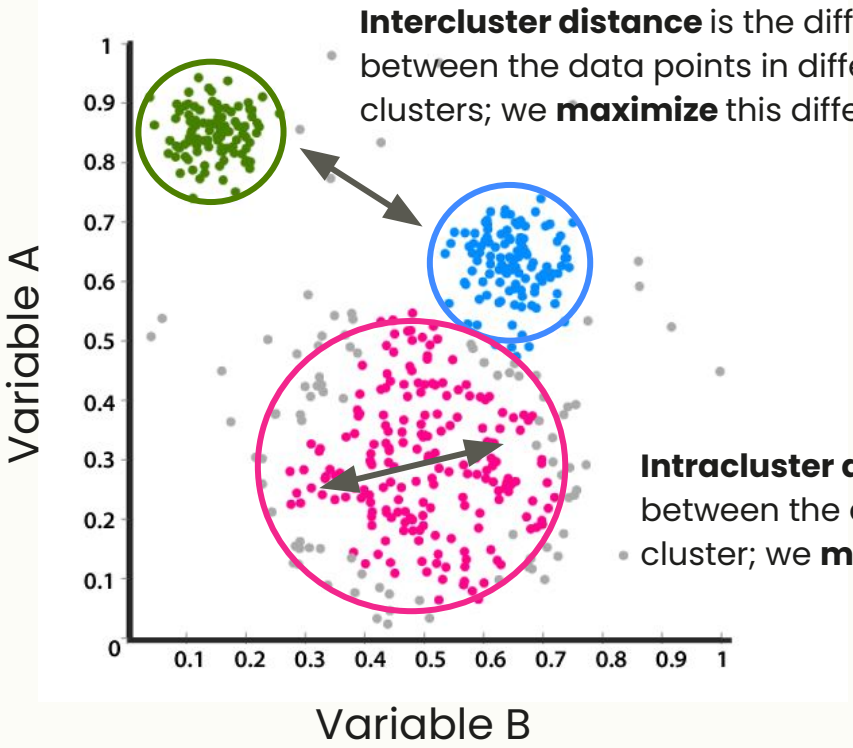


Let's say we make a scatter plot of two variables (A and B) that we've collected for 200 people

Inspection of the data suggests the 200 observations fall into 3 distinct clusters

With more observations and many more variables, we need to use statistical methods called clustering algorithms to find the clusters

There are many different methods that can be used to perform cluster analysis but all work in a similar way



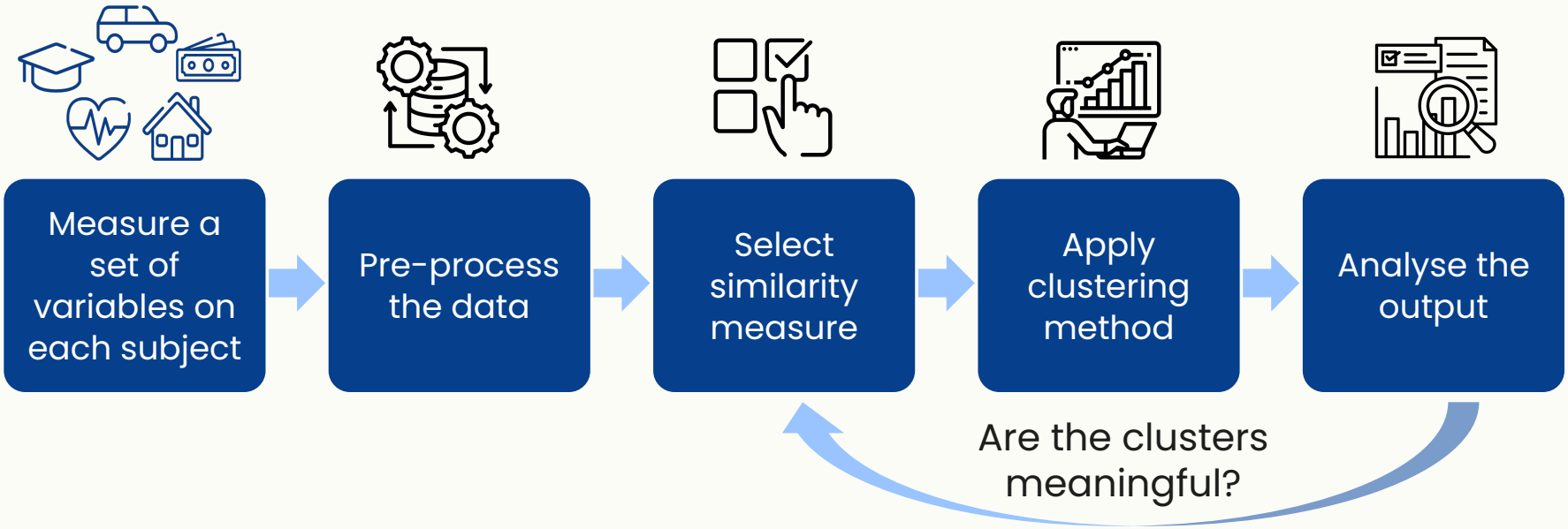
Intercluster distance is the difference between the data points in different clusters; we **maximize** this difference

Intracluster distance is the difference between the data points inside a cluster; we **minimize** this difference

Clustering measures the similarity between data points.

It groups the points that have a higher measure of similarity than points in any other cluster

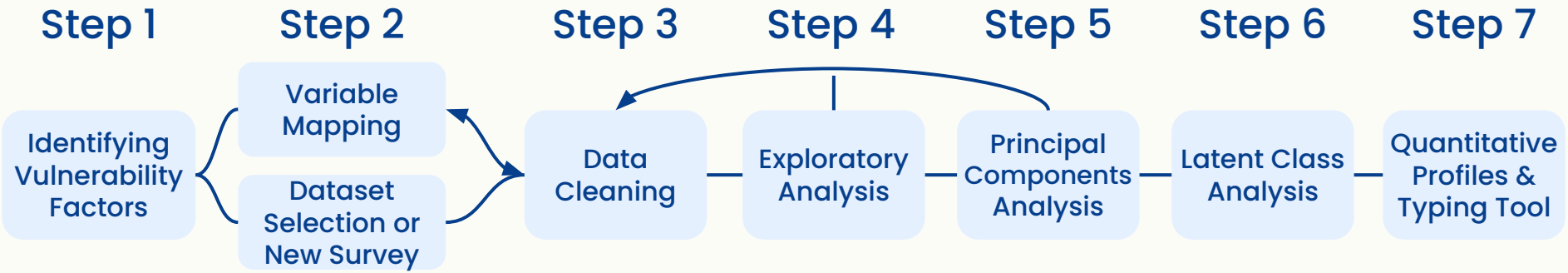
Cluster analysis requires researchers to flow through a set of steps



The **Pathways segmentation method** was designed to guide researchers through this process

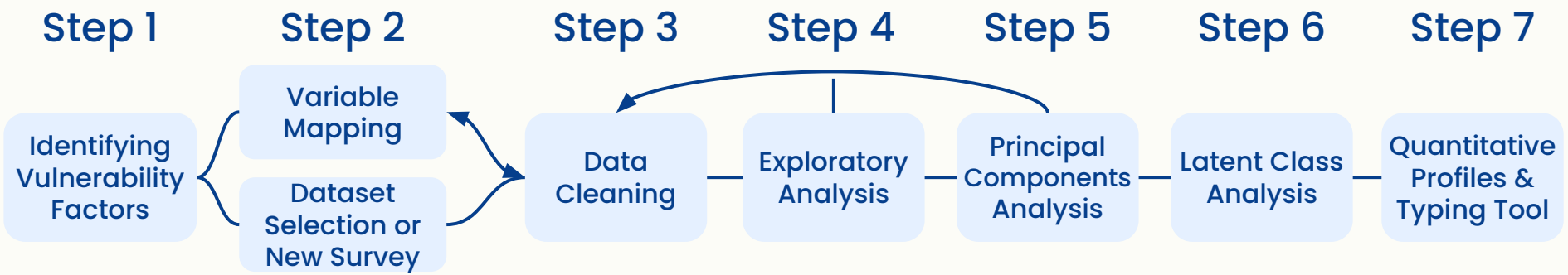
05

Pathways Segmentation Process



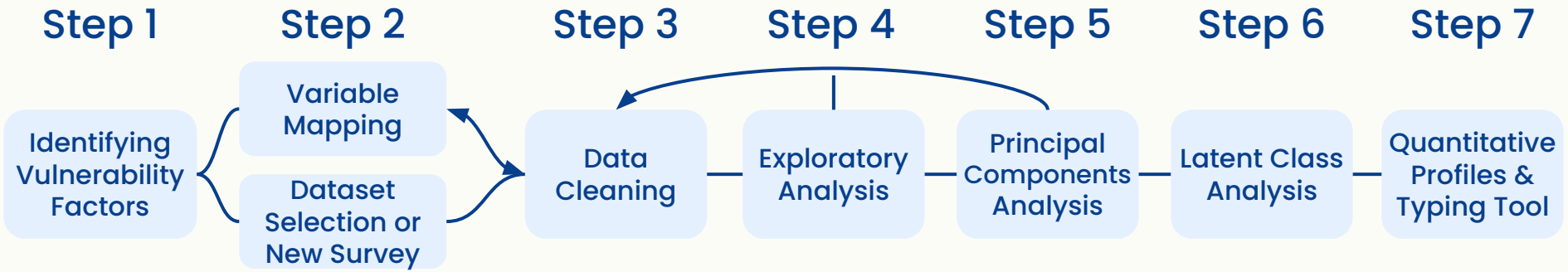
PURPOSE

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
To identify vulnerability factors for each domain that might be relevant to include.	To map variables from existing dataset(s) or new Pathways survey to the vulnerability factors.	To prepare the data for segmentation analysis.	To reduce the long list of identified variables to a smaller set that show the strongest association with behaviors/ outcomes.	To reduce the number of variables even further by removing variables that are strongly correlated (i.e. likely to measure the same thing).	To develop the segments using the variables identified in the PCA.	To describe the differences between segments using the vulnerability variables and outcome measures.



INPUTS

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
List of Pathways domains	List of vulnerability factors (by domain) List of variables in dataset	Raw dataset	Cleaned dataset	Subset of variables identified in exploratory analysis	Subset of variables identified in PCA	Segment assigned during LCA Cleaned dataset



DECISIONS

Which vulnerability factors, specific to Ethiopia, should be included

Which variables in the dataset measure each vulnerability factor

Are there factors not sufficiently measured in the data set

Can we use an existing dataset or do we need to conduct a new survey

How to group response categories for variables

What variables to combine into an index (e.g., wealth index, IPV)

Which variables to drop due to poor data quality or little variation

Alternative coding of categorical variables

What variables to include in PCA

What variables to include in LCA

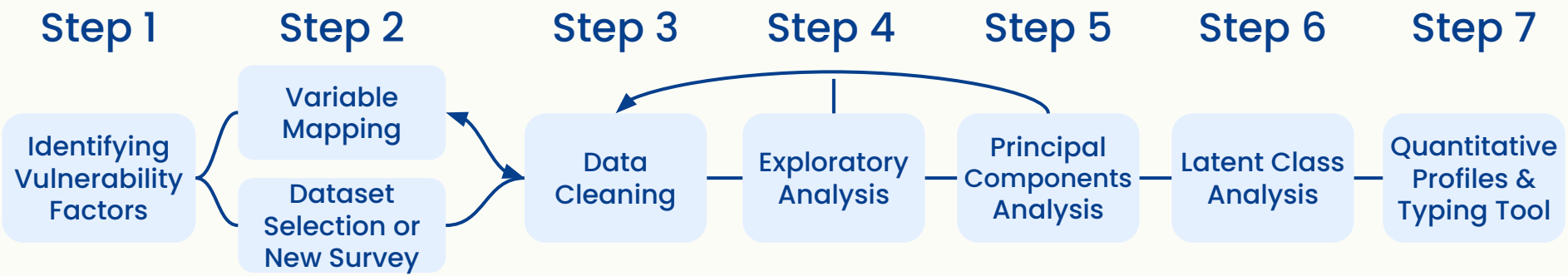
A list of variables that could be used inter-changeably to describe same vulnerability factor

How many classes or segments

What are the differentiating variables

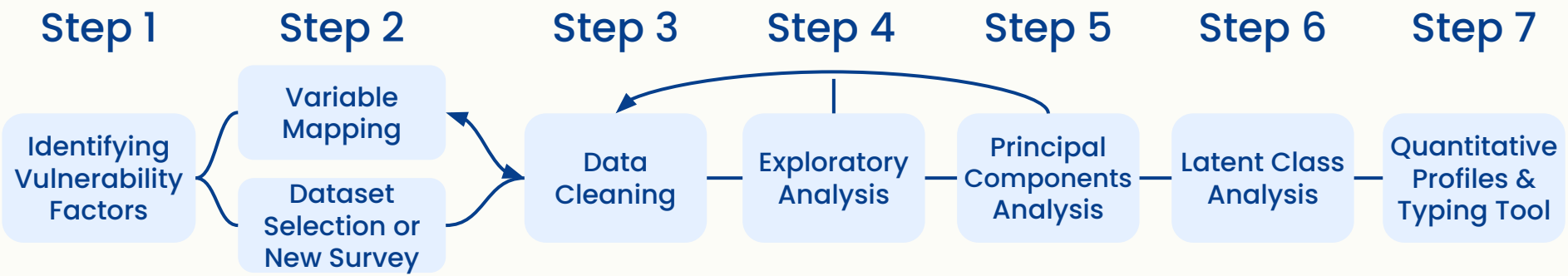
How vulnerable are women in each segment

What vulnerability factors/ outcomes help describe differences between segments



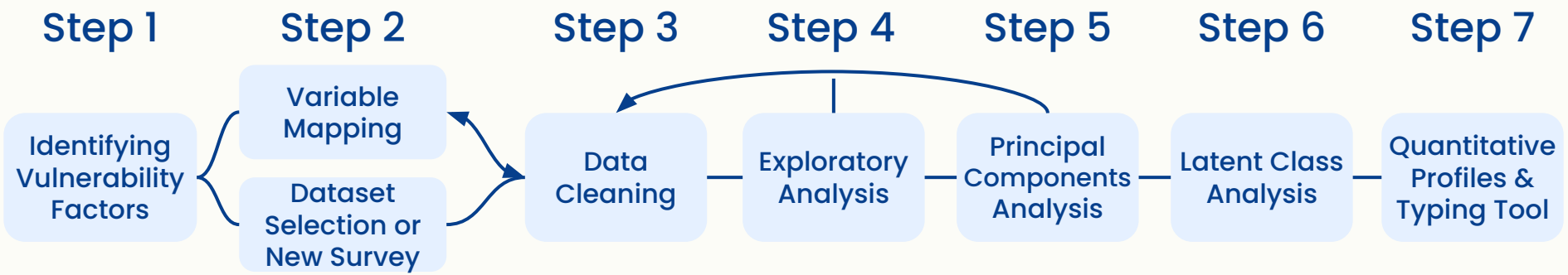
CONSIDERATIONS

Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
Sample size Availability of variables that measure vulnerability factors	Completeness and missing data Skip patterns Sample size in each category	A priori hypotheses & knowledge Sample size/ distribution Patterns across response categories Strength and no. of associations between outcomes and vulnerabilities Associations with age	Co-variability between variables by domain	Stability of segments Size of segment Model fit statistics Variables that drive the split between segments Usability concerns	Patterns in the vulnerability variables and health outcomes/ behaviors within and between segments



DECISION SUPPORT TOOLS

Step 3	Step 4	Step 5	Step 6	Step 7
Histograms	Ladder plots of predicted probabilities Histograms Tabular regression output (ORs and SEs)	Bi-plots Variable composition plot	Bar charts Sanky plots Convergence plot and related fit statistics: BIC, MLE, Posterior probability	Bar charts/ line graphs



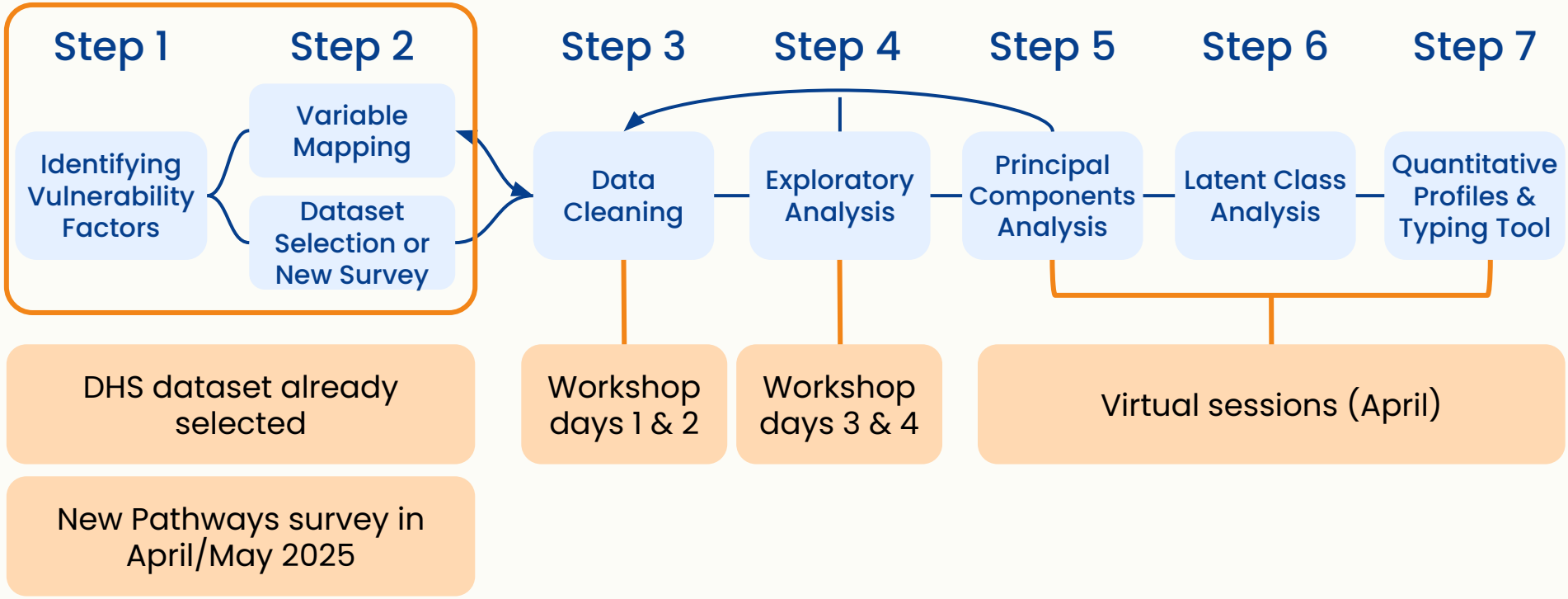
OUTPUTS

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
Spreadsheet with domains and potential factors	Spreadsheet with domains, factors, and variable names	Excel spreadsheet with initial set of coded variables Cleaned and recoded dataset	Excel spreadsheet with final list of coded variables + indicator to pass variable to PCA	Excel spreadsheet with indicator of the reduced set of variables to pass to the LCA	Number of segments + segment assignment for all records in dataset List of differentiating variables	Initial quantitative profile of each segment

discussion

Which part of the process might be most challenging for you and why?

Timeline for segmentation training



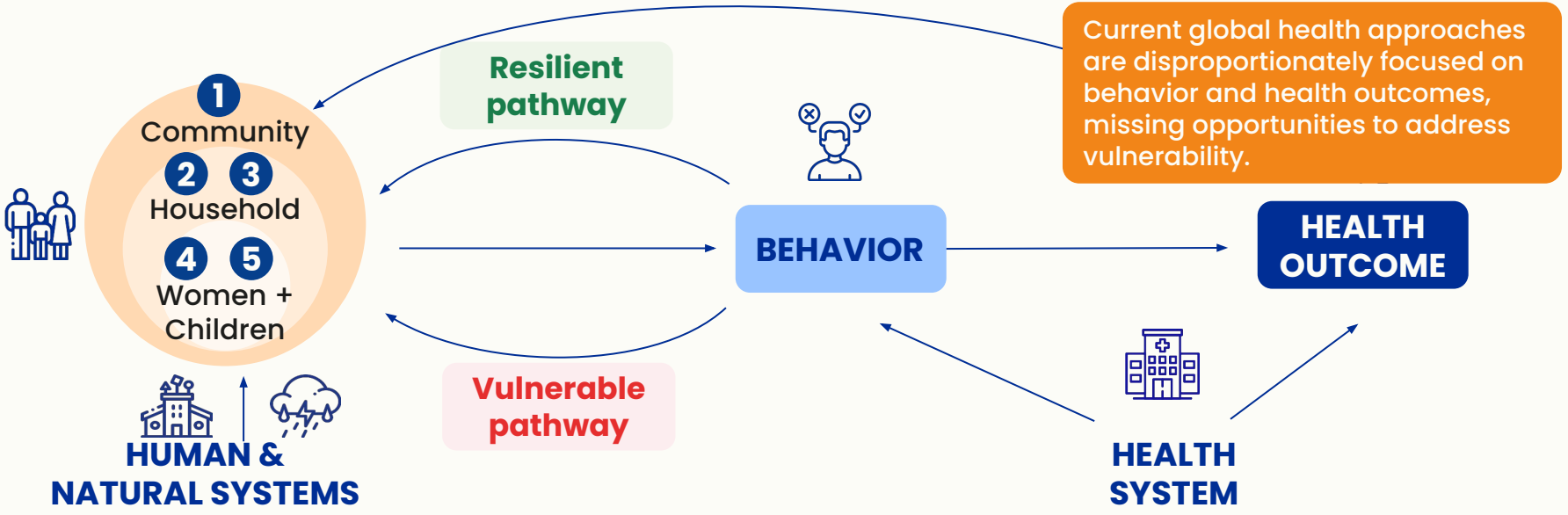
06

Identifying Vulnerability Factors

Vulnerability conceptual framework

Vulnerability domains:

1. Social support
2. Household economics
3. Household relationships & living conditions
4. Health Mental Models
5. Woman and her past experience
6. Human & natural systems



The woman & her past experience

Household relationships

Household Economics & Living conditions

Social Support

Health mental models

Human & natural systems

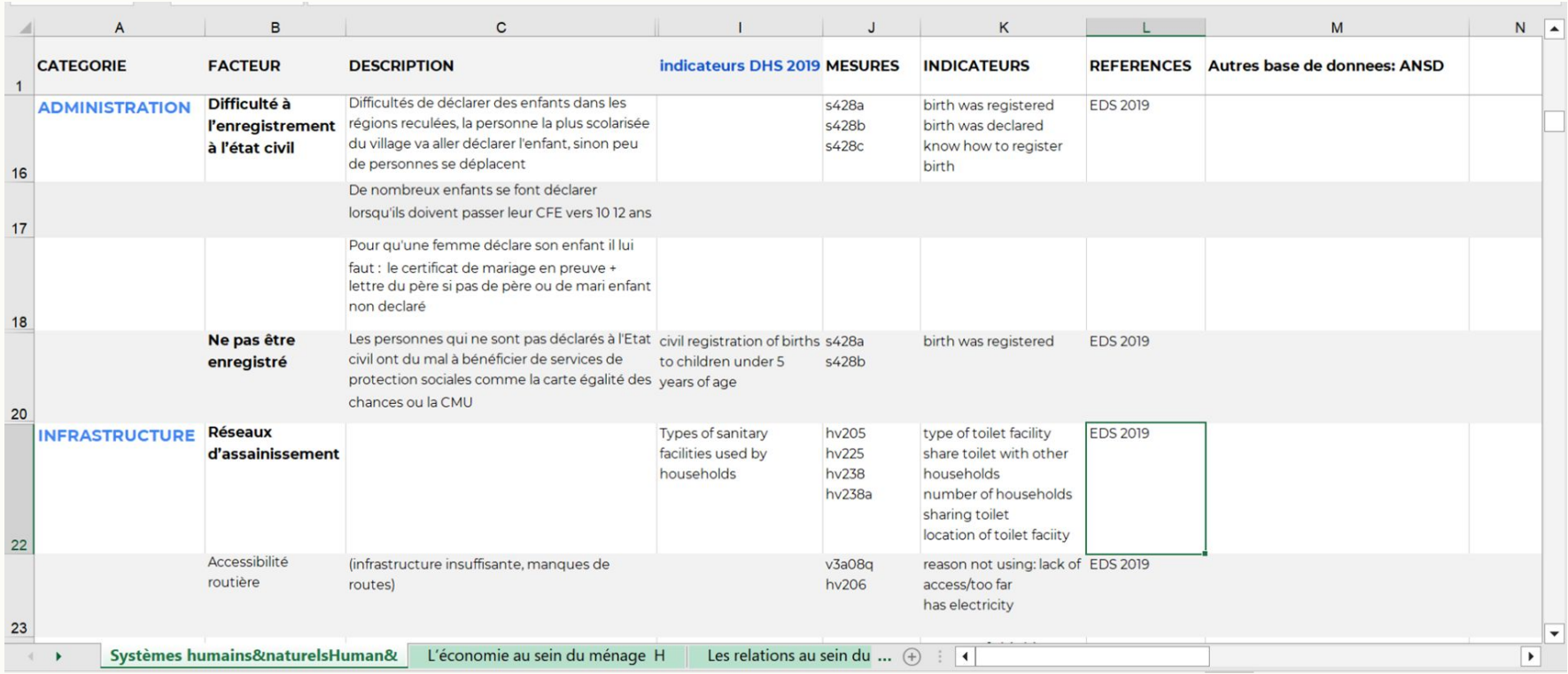
Variable mapping to Pathways

DOMAIN	FACTOR	SURVEY QUESTION	
Woman and Her Past Experience	Woman's education level	What is the highest level of school you attended? Question relevant when: ${consent_obtained}	[0] Never attended [1] Primary [2] Secondary [3] Higher [4] No response
Woman and Her Past Experience; Household Relationships	Matrital status	What is your current martital status?	[0] Never in union [1] Married [2] Living with partner [3] Widowed [4] Divorced [5] No longer living together/separated
Household Relationships	Decision-making on child's health	Who do you consult with for decisions regarding your children's healthcare?	1. No one, I decide on my own 2. My husband. 3. Mother-in-law 4. Father-in-law 5. Relatives from my husband's side -96. Other (specify): _____ -98. Don't know -99. Refused to answer
Household Economics and Living Conditions	Woman's ownership of bank account	Do you have a bank account? [Do not read list]	[0] No I don't [1] Has own bank account alone [2] Has joint account [3] Has both joint and separate accounts
Household Economics and Living Conditions	Woman's ownership of mobile money account	Do you have a mobile banking app or account {country-specific, such as Paytm, M-Pesa, etc. } [Do not read list]	[0] No I don't [1] Has own account [2] Has joint account [3] Has both joint and separate accounts

Variable mapping to Pathways

DOMAIN	FACTOR	SURVEY QUESTION	
Social Support	Size of support	How many people are so close to you that you can count on them if you have great personal problems?	'none' '1–2' '3–5' '5+'
Social Support		How easy is it to get practical help from neighbors should you need it?	very difficult' 'difficult' 'possible' 'easy' 'very easy'
Health Mental Model	C03	Had you heard about sex before you had your first sexual experience?	Yes/No
Health Mental Model	C01	Had you heard about menstruation before you got your first period?	Yes/No
Human and Natural System	Perceived impact of climate	Now I'm going to read a series of statements about the weather. Please answer yes or no after each one. [PROBE: even if you aren't sure, your best guess is fine]	2 COLUMNS: [1] Yes [2] No {Randomize} ROWS: [1] People in my community are more sick today because of dirty air [2] People in my community are more sick today because of dirty water [3] My community has faced more drought or floods [4] It is harder to predict when there will be rain

Variable mapping to DHS

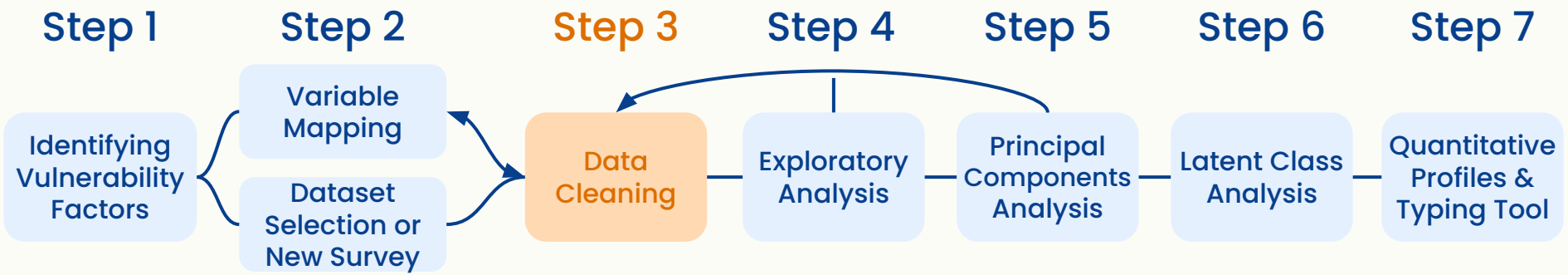


	A	B	C	I	J	K	L	M	N
1	CATEGORIE	FACTEUR	DESCRIPTION	indicateurs DHS 2019	MESURES	INDICATEURS	REFERENCES	Autres base de donnees: ANSD	
16	ADMINISTRATION	**Difficulté à l'enregistrement à l'état civil**	Difficultés de déclarer des enfants dans les régions reculées, la personne la plus scolarisée du village va aller déclarer l'enfant, sinon peu de personnes se déplacent		s428a s428b s428c	birth was registered birth was declared know how to register birth	EDS 2019		
17			De nombreux enfants se font déclarer lorsqu'ils doivent passer leur CFE vers 10 12 ans						
18			Pour qu'une femme déclare son enfant il lui faut : le certificat de mariage en preuve + lettre du père si pas de père ou de mari enfant non declaré						
20		**Ne pas être enregistré**	Les personnes qui ne sont pas déclarés à l'Etat civil ont du mal à bénéficier de services de protection sociales comme la carte égalité des chances ou la CMU	civil registration of births to children under 5 years of age	s428a s428b	birth was registered	EDS 2019		
22	INFRASTRUCTURE	**Réseaux d'assainissement**		Types of sanitary facilities used by households	hv205 hv225 hv238 hv238a	type of toilet facility share toilet with other households number of households sharing toilet location of toilet faciity	EDS 2019		
23		Accessibilité routière	(infrastructure insuffisante, manques de routes)		v3a08q hv206	reason not using: lack of access/too far has electricity	EDS 2019		

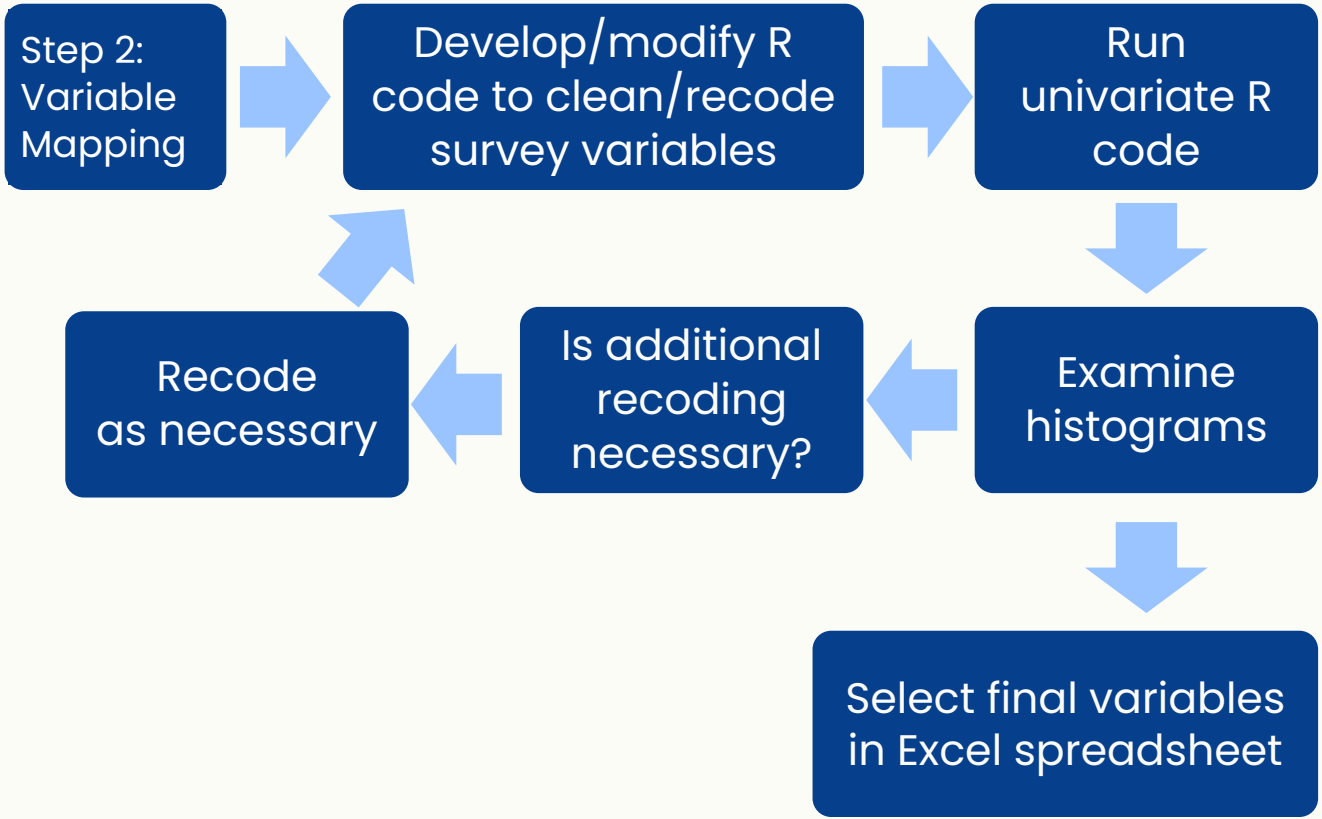
Systèmes humains&naturelsHuman& | L'économie au sein du ménage H | Les relations au sein du ...

07

Data Cleaning


Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Identifying Vulnerability Factors
Variable Mapping
Dataset Selection or New Survey
Data Cleaning
Exploratory Analysis
Principal Components Analysis
Latent Class Analysis
Quantitative Profiles & Typing Tool

Data cleaning and variable generation process



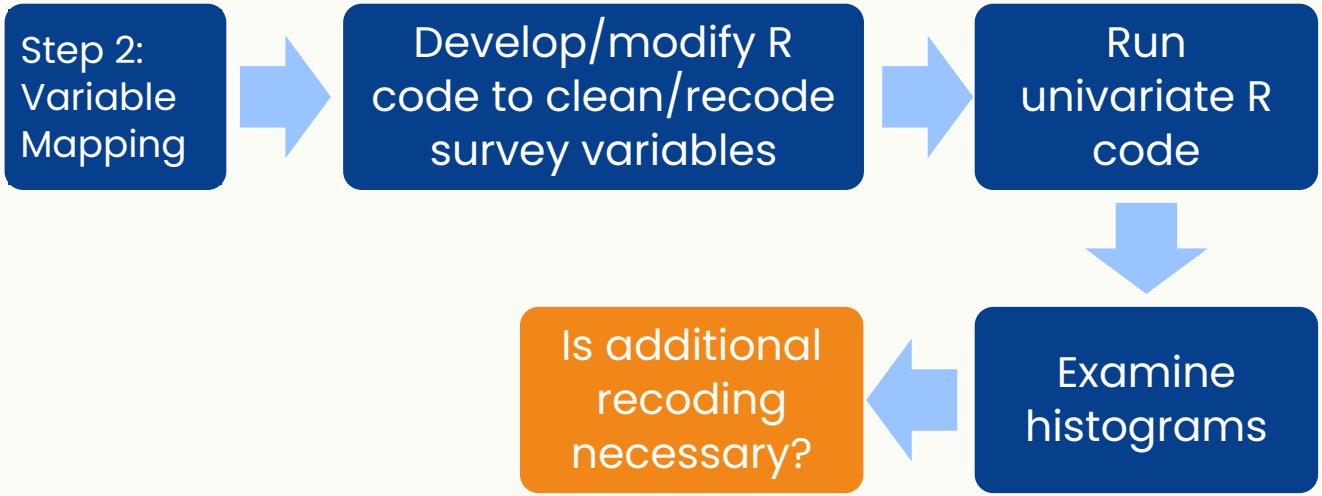
Data cleaning and variable generation process

Step 2: Variable Mapping → Develop/modify R code to clean/recode survey variables

Questions to consider in this step:

- Do we need to recode variables to improve data quality?
- Should we group multiple response categories?
- Do we combine multiple variables to create an index?
- Are there continuous variables that should be recoded to categorical variables?

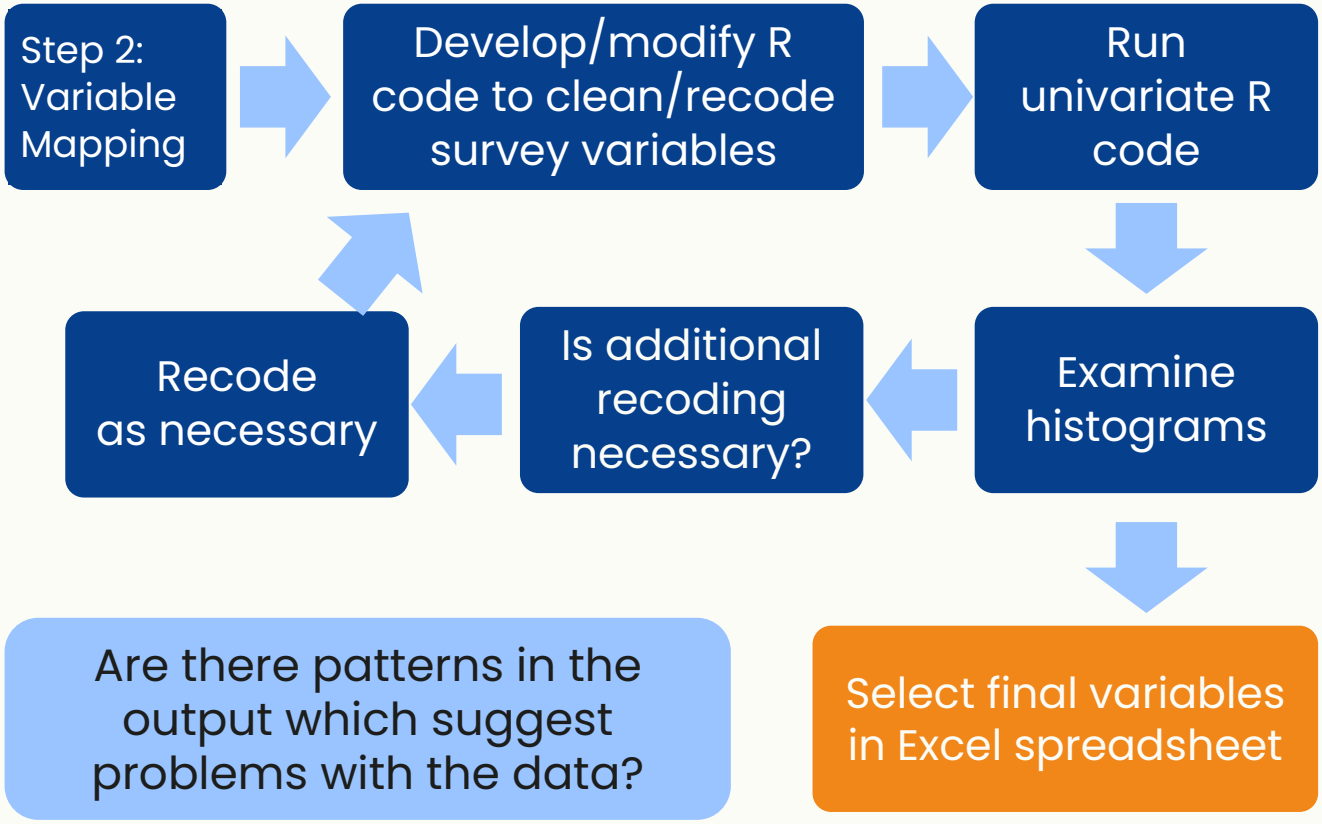
Data cleaning and variable generation process



Questions to consider in this step:

- What is the sample size/prop. of sample in each response category?
- Are there a large number of missing values?
  - Is this a result of survey skip patterns?
- Is the variable related to respondent age?

Data cleaning and variable generation process



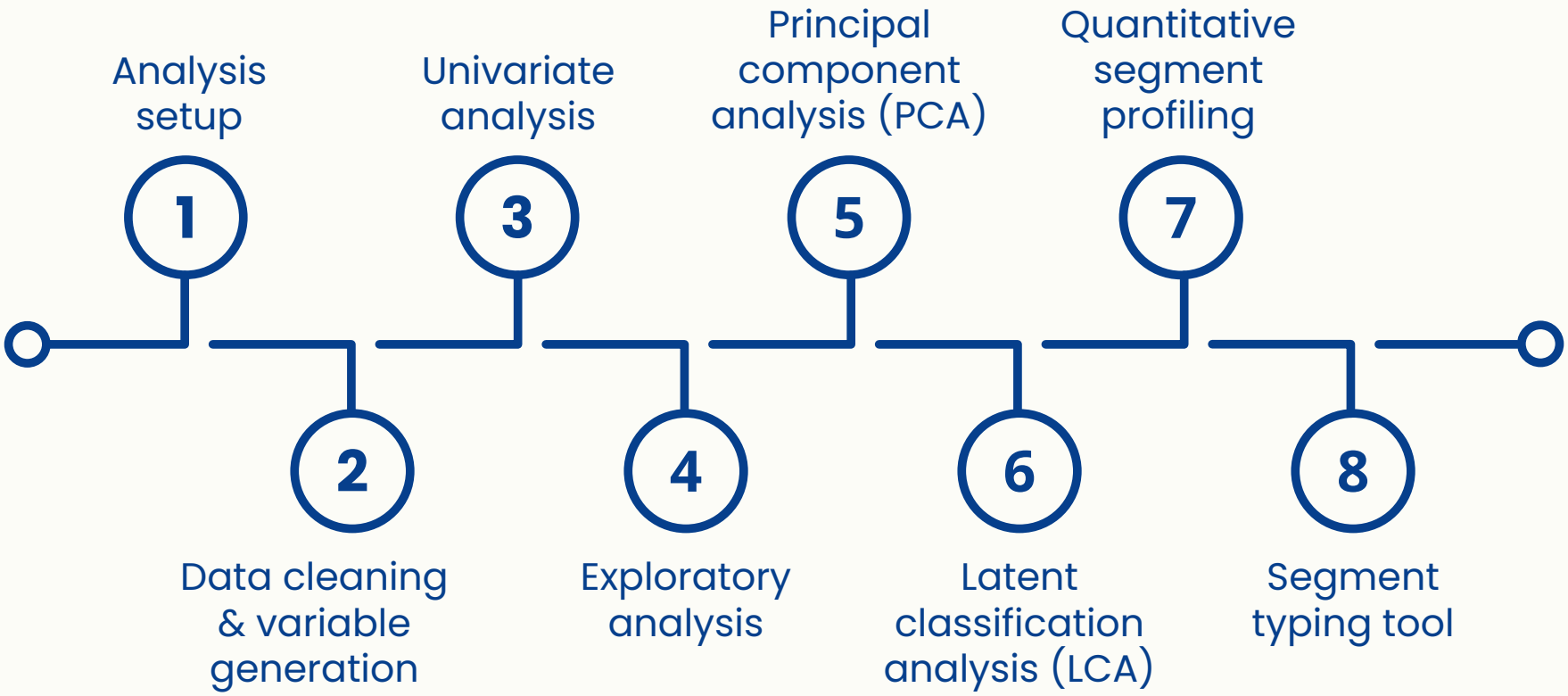
Consider dropping a variable if:

- There is no variation across response categories
- There is a large number of missing values

08

Technical Guide

Analysis stages



Workflow technical requirements

Microsoft

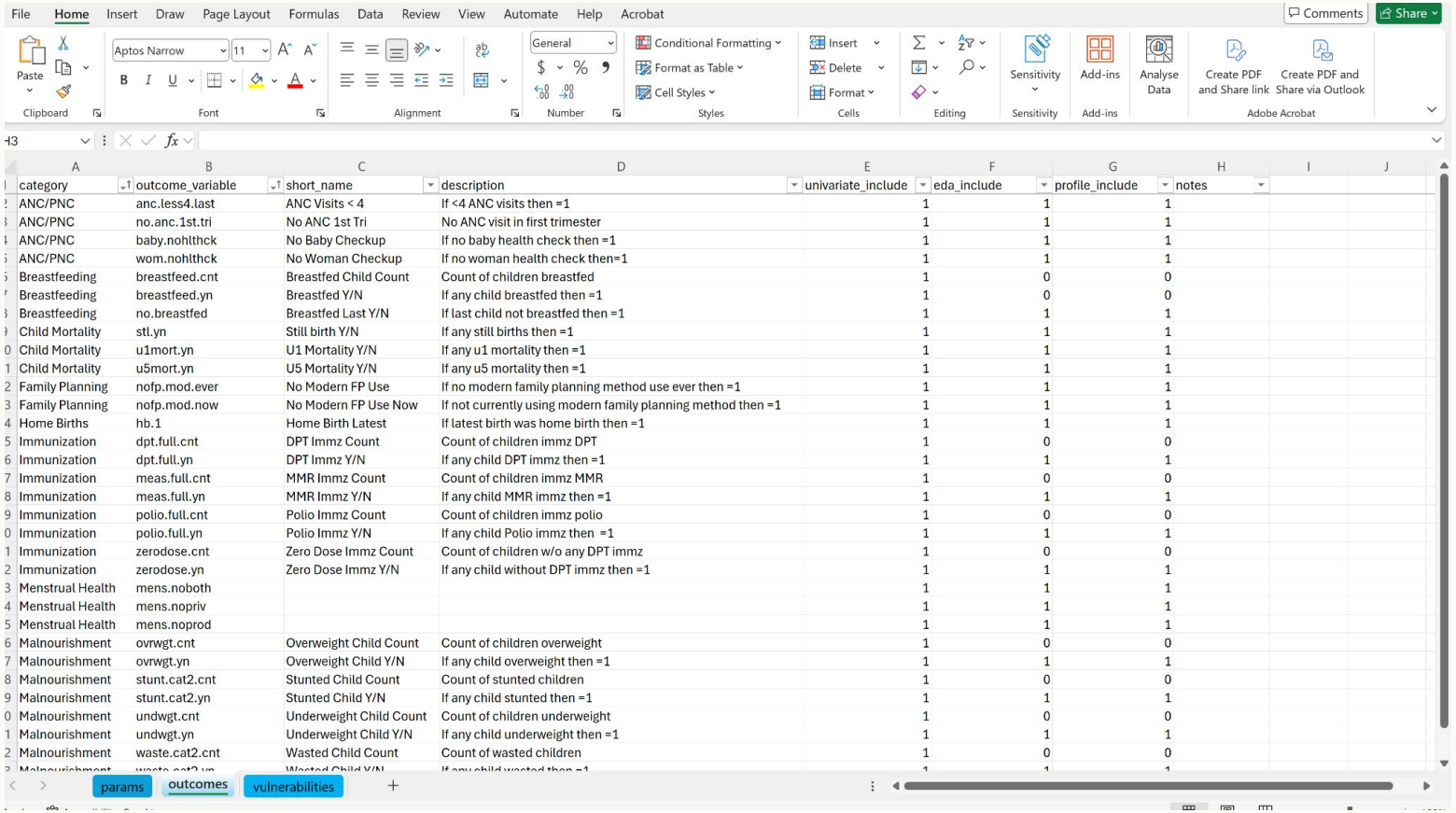
GitHub

(suggest GitHub Desktop)

R Studio

(+R 4.4.2 recommended)

Pathways workbook



category	outcome_variable	short_name	description	univariate_include	eda_include	profile_include	notes
ANC/PNC	anc.less4.last	ANC Visits < 4	If <4 ANC visits then =1	1	1	1	
ANC/PNC	no.anc.1st.tri	No ANC 1st Tri	No ANC visit in first trimester	1	1	1	
ANC/PNC	baby.nohlthck	No Baby Checkup	If no baby health check then =1	1	1	1	
ANC/PNC	wom.nohlthck	No Woman Checkup	If no woman health check then=1	1	1	1	
Breastfeeding	breastfeed.cnt	Breastfed Child Count	Count of children breastfed	1	0	0	
Breastfeeding	breastfeed.yn	Breastfed Y/N	If any child breastfed then =1	1	0	0	
Breastfeeding	no.breastfed	Breastfed Last Y/N	If last child not breastfed then =1	1	1	1	
Child Mortality	stl.yn	Still birth Y/N	If any still births then =1	1	1	1	
Child Mortality	u1mort.yn	U1 Mortality Y/N	If any u1 mortality then =1	1	1	1	
Child Mortality	u5mort.yn	U5 Mortality Y/N	If any u5 mortality then =1	1	1	1	
Family Planning	nofp.mod.ever	No Modern FP Use	If no modern family planning method use ever then =1	1	1	1	
Family Planning	nofp.mod.now	No Modern FP Use Now	If not currently using modern family planning method then =1	1	1	1	
Home Births	hb.1	Home Birth Latest	If latest birth was home birth then =1	1	1	1	
Immunization	dpt.full.cnt	DPT Immz Count	Count of children immz DPT	1	0	0	
Immunization	dpt.full.yn	DPT Immz Y/N	If any child DPT immz then =1	1	1	1	
Immunization	meas.full.cnt	MMR Immz Count	Count of children immz MMR	1	0	0	
Immunization	meas.full.yn	MMR Immz Y/N	If any child MMR immz then =1	1	1	1	
Immunization	polio.full.cnt	Polio Immz Count	Count of children immz polio	1	0	0	
Immunization	polio.full.yn	Polio Immz Y/N	If any child Polio immz then =1	1	1	1	
Immunization	zerodose.cnt	Zero Dose Immz Count	Count of children w/o any DPT immz	1	0	0	
Immunization	zerodose.yn	Zero Dose Immz Y/N	If any child without DPT immz then =1	1	1	1	
Menstrual Health	mens.noboth			1	1	1	
Menstrual Health	mens.nopriv			1	1	1	
Menstrual Health	mens.noprod			1	1	1	
Malnourishment	ovrwgt.cnt	Overweight Child Count	Count of children overweight	1	0	0	
Malnourishment	ovrwgt.yn	Overweight Child Y/N	If any child overweight then =1	1	1	1	
Malnourishment	stunt.cat2.cnt	Stunted Child Count	Count of stunted children	1	0	0	
Malnourishment	stunt.cat2.yn	Stunted Child Y/N	If any child stunted then =1	1	1	1	
Malnourishment	undwgt.cnt	Underweight Child Count	Count of children underweight	1	0	0	
Malnourishment	undwgt.yn	Underweight Child Y/N	If any child underweight then =1	1	1	1	
Malnourishment	waste.cat2.cnt	Wasted Child Count	Count of wasted children	1	0	0	

Pathways workbook

27	Malnourishment	ovrwgt.yn	Overweight Child Y/N	If any child overweight then =1
28	Malnourishment	stunt.cat2.cnt	Stunted Child Count	Count of stunted children
29	Malnourishment	stunt.cat2.yn	Stunted Child Y/N	If any child stunted then =1
30	Malnourishment	undwgt.cnt	Underweight Child Count	Count of children underweight
31	Malnourishment	undwgt.yn	Underweight Child Y/N	If any child underweight then =1
32	Malnourishment	waste.cat2.cnt	Wasted Child Count	Count of wasted children
33	Malnourishment	waste.cat2.yn	Wasted Child Y/N	If any child wasted then =1

params | outcomes | vulnerabilities

Params: columns for strata, domains, etc. Serves as a catch-all for some project-specific variables

Pathways workbook

27	Malnourishment	ovrwgt.yn	Overweight Child Y/N	If any child overweight then =1
28	Malnourishment	stunt.cat2.cnt	Stunted Child Count	Count of stunted children
29	Malnourishment	stunt.cat2.yn	Stunted Child Y/N	If any child stunted then =1
30	Malnourishment	undwgt.cnt	Underweight Child Count	Count of children underweight
31	Malnourishment	undwgt.yn	Underweight Child Y/N	If any child underweight then =1
32	Malnourishment	waste.cat2.cnt	Wasted Child Count	Count of wasted children
33	Malnourishment	waste.cat2.yn	Wasted Child Y/N	If any child wasted then =1

params | outcomes | vulnerabilities

Outcomes: Health behaviors/outcomes and metadata used in the analysis; inclusion/exclusion decisions for each phase are driven from this tab.

Pathways workbook

27	Malnourishment	ovrwgt.yn	Overweight Child Y/N	If any child overweight then =1
28	Malnourishment	stunt.cat2.cnt	Stunted Child Count	Count of stunted children
29	Malnourishment	stunt.cat2.yn	Stunted Child Y/N	If any child stunted then =1
30	Malnourishment	undwgt.cnt	Underweight Child Count	Count of children underweight
31	Malnourishment	undwgt.yn	Underweight Child Y/N	If any child underweight then =1
32	Malnourishment	waste.cat2.cnt	Wasted Child Count	Count of wasted children
33	Malnourishment	waste.cat2.yn	Wasted Child Y/N	If any child wasted then =1

params | outcomes | vulnerabilities

Vulnerabilities: Vulnerability variables and metadata used in the analysis; inclusion/exclusion decisions for each phase are driven from this tab.

1 Analysis setup

Pathways Workbook Columns

N/A

Workflow R Scripts

1_setup.R

Outputs

Pathways Data Dictionary

Set up R environment

1. Install/load all required R libraries using the **renv** framework
2. Read in variables defined in the **config.yml** file
3. Create folder structure for the analysis and define file paths
4. Import Pathways Data Dictionary and Pathways Workbook (if applicable)

2 Data cleaning and variable generation

Pathways Workbook Columns

N/A

Workflow R Scripts

2_data_cleaning.R
functions/fun_gen_vulnerabilities.R
functions/fun_gen_outcomes.R

Outputs

Outcomes (df)
Vulnerability (df)
Outcomes_vulnerability (df)
Pathways Workbook (xlsx)

Coding variable guidance

- Health behaviors/outcomes should be defined as binary 0/1 variables with 1 representing the less desirable health/behavioral state
- Vulnerability variables should be coded as binary/ categorical data type
- Exceptions can be made for the exploratory phase, but these criteria should be met before beginning the Principal Component Analysis phase

3 Univariate analysis

Pathways Workbook Columns

outcomes.univariate_include

Vulnerability.univariate_include

Workflow R Scripts

3_univariate_analysis.R

functions/fun_univariate_visuals.R

Outputs

Univariate_plots_outcomes.pdf

Univariate_plots_vulnerability.pdf

Univariate plots help us to understand

- Is the variable coded as expected?
- What is the sample size of the variable?
- Can categories be collapsed/combined to simplify the variable?
- If we coded a variable as numeric, is there a clear shift in the distribution where we can make it binary/categorical?
- Are data types correct?

4 Exploratory analysis

Pathways Workbook Columns

Outcomes.eda_include

Vulnerability.eda_include

Workflow R Scripts

4_exploratory_data_analysis.R

functions/fun_eda.R

Outputs

eda_{x}.pdf (where x is the vulnerability variable)

4 Exploratory analysis

Pathways Workbook Columns

Outcomes.eda_include

Vulnerability.eda_include

Workflow R Scripts

4_exploratory_data_analysis.R

functions/fun_eda.R

Outputs

eda_{x}.pdf (where x is the vulnerability variable)

EDA output contains bivariate regression results by vulnerability variable and the full set of health outcomes/ behaviors

The EDA outputs help us determine

- Is the vulnerability variable (e.g., Any Media Consumption) consistently associated with important health outcomes/behaviors (e.g., ANC visits, U5 mortality)
- Can some levels in a categorical variable be collapsed based on their relationship with important health outcomes/behaviors?

    i.e., is the predicted probability of the health outcome/behavior effectively the same for multiple categories?

- Do some relationships exist that are contrary to what we believe? This could be motivation to check the variable coding.
- Are data types correct?

5 Principal component analysis (PCA)

Pathways Workbook Columns

Vulnerability.pca_strata

Vulnerability.pca_include

Workflow R Scripts

5_principal_component_analysis.R

functions/fun_pca_output.R

Outputs

pca_plots_{x}.pdf (where x is the segmentation strata)

5 Principal component analysis (PCA)

Pathways Workbook Columns

Vulnerability.pca_strata

Vulnerability.pca_include

Workflow R Scripts

5_principal_component_analysis.R

functions/fun_pca_output.R

Outputs

pca_plots_{x}.pdf (where x is the Segmentation strata)

Variables defined by setting the pca_strata appropriately and the pca_include column to 1 in the vulnerability tab in the Pathways Workbook.

The PCA outputs help us determine

- Install/load all required R libraries using the renv framework
- Read in variables defined in the config.yml file
- Create folder structure for the analysis and define file paths
- Import data dictionary and Pathways Workbook (if applicable)

6 Latent classification analysis (LCA)

Pathways Workbook Columns

vulnerability.lca_strata

vulnerability.lca_include

Workflow R Scripts

6_latent_classification_analysis.R

functions/fun_lca_output.R

functions/fun_lca_output_visuals.R

functions/fun_lca_exploratory.R

Outputs

{x}_outcomes_vulnerability_class (df)

{x}_lca_output_plots.pdf

{x}_lca_exploratory_plots.pdf

(where x is segmentation strata)

6 Latent classification analysis (LCA)

Pathways Workbook Columns

vulnerability.lca_strata
vulnerability.lca_include

Workflow R Scripts

6_latent_classification_analysis.R
functions/fun_lca_output.R
functions/fun_lca_output_visuals.R
functions/fun_lca_exploratory.R

Outputs

{x}_outcomes_vulnerability_class (df)
{x}_lca_output_plots.pdf
{x}_lca_exploratory_plots.pdf
(where x is Segmentation strata)

Variables defined by setting the pca_strata appropriately and the pca_include column to 1 in the vulnerability tab in the Pathways Workbook.

6 Latent classification analysis (LCA)

Pathways Workbook Columns

vulnerability.lca_strata

vulnerability.lca_include

Workflow R Scripts

6_latent_classification_analysis.R

functions/fun_lca_output.R

functions/fun_lca_output_visuals.R

functions/fun_lca_exploratory.R

Outputs

{x}_outcomes_vulnerability_class (df)

{x}_lca_output_plots.pdf

{x}_lca_exploratory_plots.pdf
(where x is segmentation strata)

LCA outputs:
Diagnostic plots (statistical properties of classification solution)
Exploratory plots (differences across the input variables based on the classes of the output)

7 Quantitative segment profiling

Pathways Workbook Columns

vulnerability.profile_include
outcomes.profile_include
params.final_model

Workflow R Scripts

7_segment_profiles.R

functions/fun_segment_profiles.R

Outputs

{x}_segment_profile_plots.pdf (where x is segmentation strata)

8 Segment typing tool

Pathways Workbook Columns

vulnerability.typing_tool_strata

vulnerability.typing_tool_include

Workflow R Scripts

8_typing_tool.R

functions/fun_typing_tool.R

Outputs

{x}_typing_tool_plots.pdf (where x is segmentation strata)

activity

Let’s take a look at the Pathways Workbook in Excel and R code for data cleaning.

activity

Setting up the R environment and running the segmentation code

activity

Implementing the data cleaning and univariate code (in R)

https://forms.gle/Hzw3DnxJzPnPujmM9

End-of-day survey

pathways

Pathways Segmentation Methods Workshop

Day 2 – Data Cleaning & Recoding

IDM Gates Foundation

Day 2 Outline

- Review from day 1
- Examples of data cleaning
- Activity: Clean and recode Ethiopia DHS data
- Activity: Consensus on variables
- Office hours

01

Day 1 Review

discussion

What did you learn from yesterday?

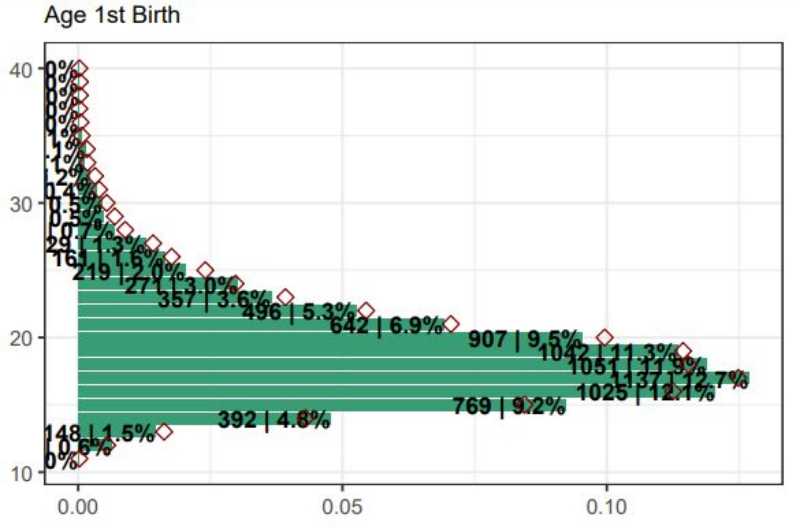
Was there anything confusing or that you would like to review?

02

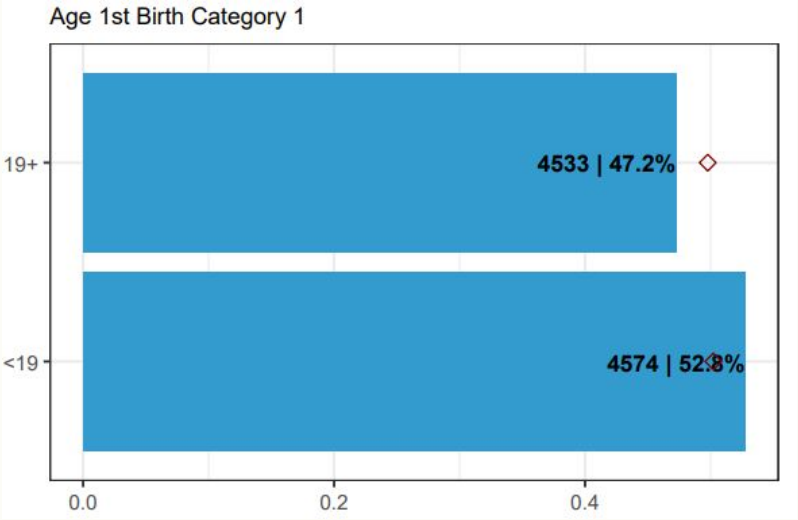
Examples of Data Cleaning

Create a binary categorical variable from a continuous variable

Example of a continuous variable

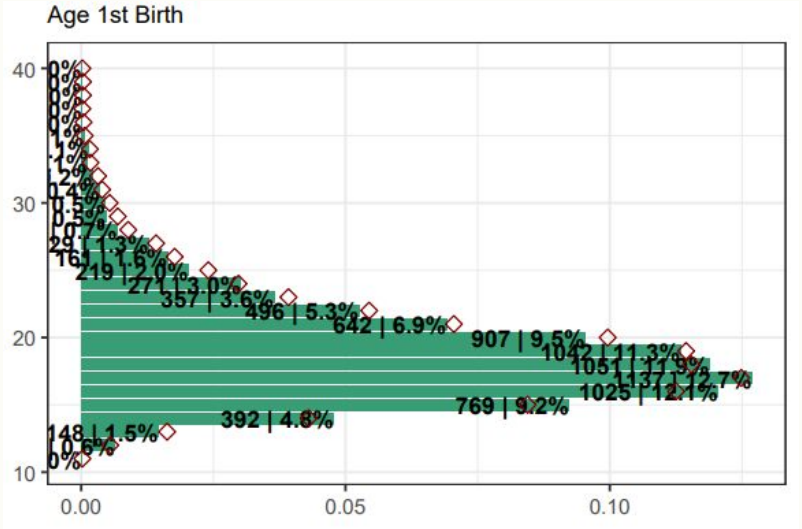


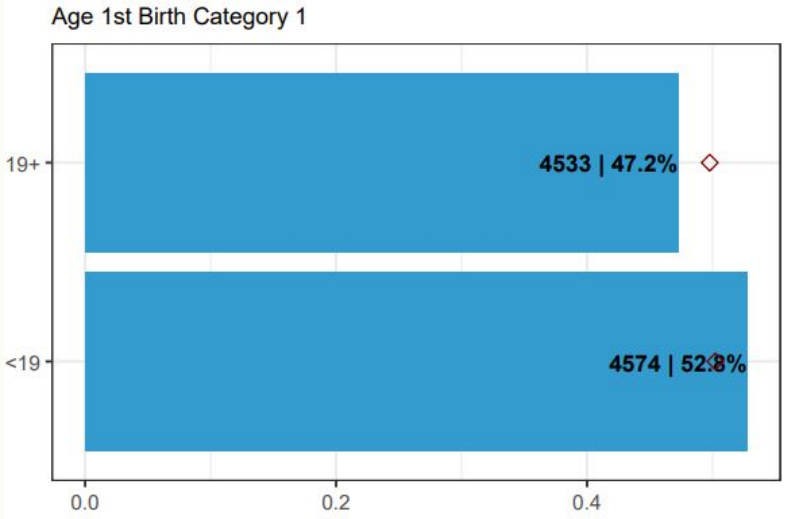
Example of binary recoding



Where should we split the data to create a new categorical variable?

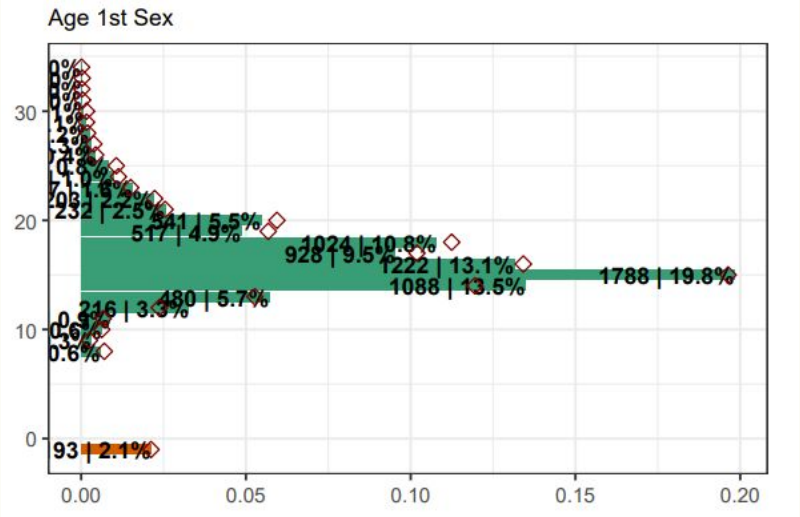
Create a binary categorical variable from a continuous variable

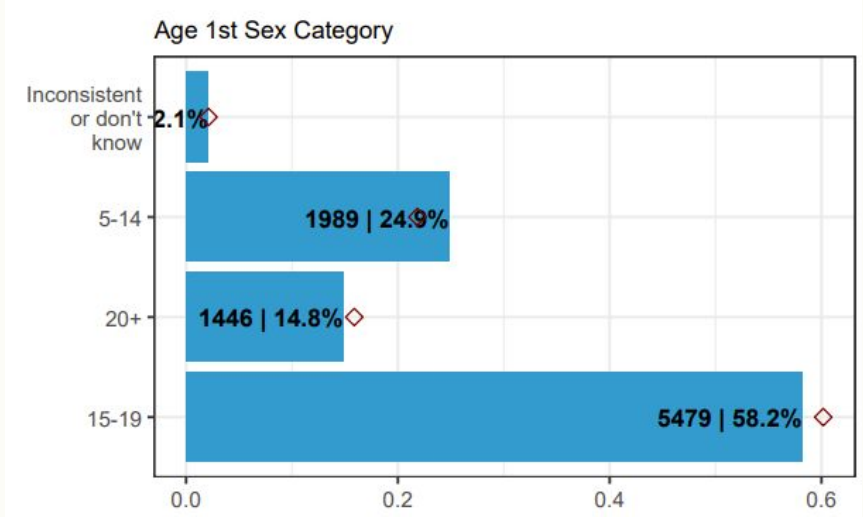





```
####################################
# CREATE BINARY CATEGORICAL VARIABLE BY SPLITTING THE DISTRIBUTION INTO TWO GROUPS
# PRESERVE NA VALUES AS NA
df <- df %>%
  dplyr::mutate(age.1stbrth.cat = case_when(age.1stbrth >= 19 ~ "19+",
                                            age.1stbrth < 19 ~ "<19"))
```

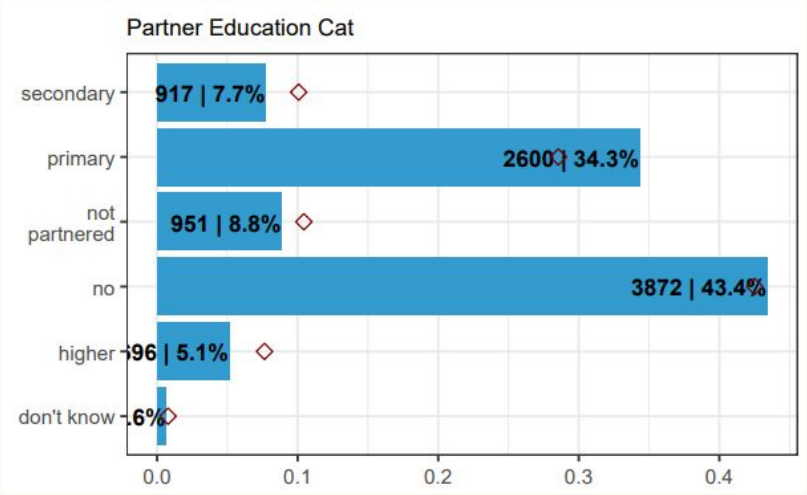
Create a multi-category variable from a continuous variable

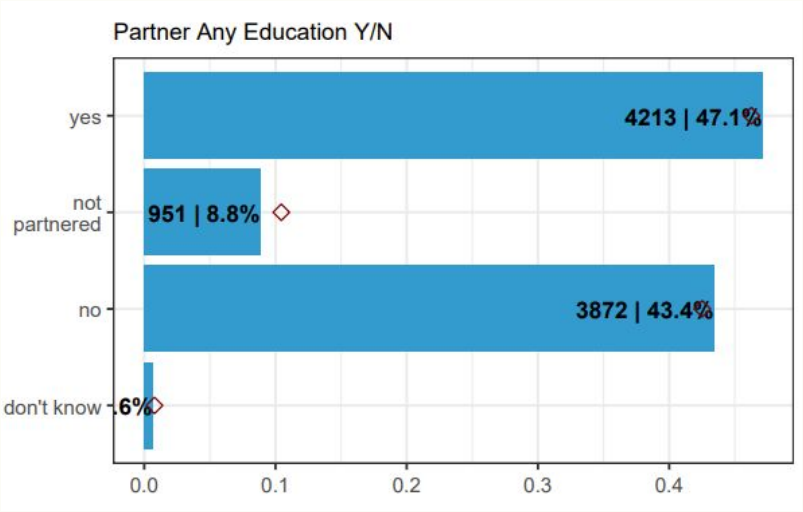





```
####################################
# CATEGORICAL FACTOR FOR AGE AT FIRST SEX
# ACCOUNTING FOR SKIP PATTERNS WHICH CREATE NA
IR<- IR %>%
  dplyr::mutate(age.1stsex.cat = case_when(v531 %in% c(0, "not had sex") ~ "never",
                                           (age.1stsex > 0 & age.1stsex < 15) ~ "5-14",
                                           (age.1stsex >= 15 & age.1stsex < 20) ~ "15-19",
                                           (age.1stsex >= 20 & age.1stsex < 50) ~ "20+",
                                           v531 %in% c(97, 98, "inconsistent", "don't know") ~ "Inconsistent or don't know"))
```

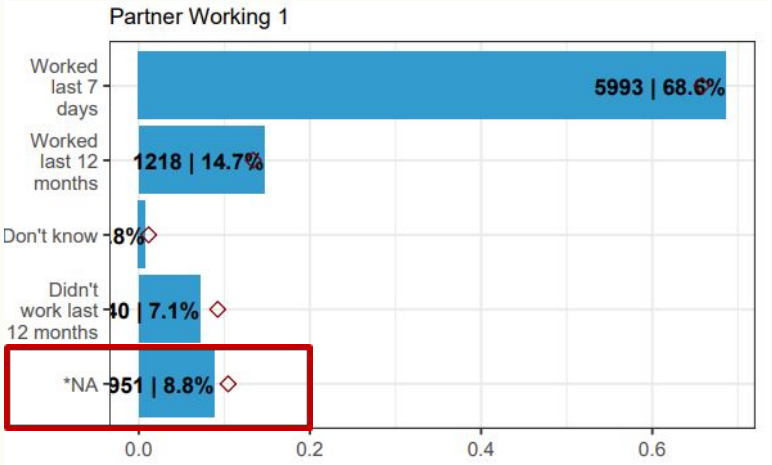
Collapse categories of a variable to create a new simplified variable

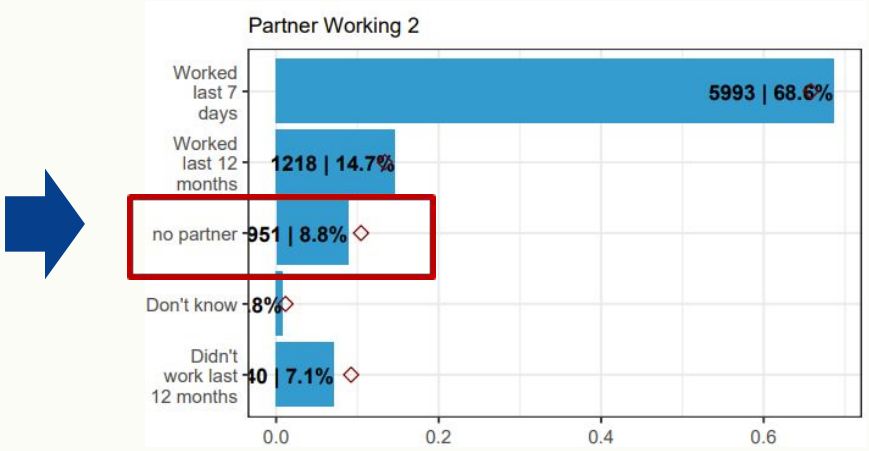





```
###################################
IR <- IR %>% dplyr::mutate(partner.anyed.yn = case_when(
  partner.ed.level == "no education" ~ "no",
  partner.ed.level %in% c("higher","primary","secondary") ~ "yes",
  partner.ed.level == "don't know" ~ "don't know",
  partner.ed.level == "not partnered" ~ "not partnered"))
```

Account for survey skip patterns when creating a categorical variable






```
####################################
# HUSBAND/PARTNER WORKING 2
# ACCOUNT FOR PARTNER STATUS SURVEY SKIP PATTERN
IR <- IR %>% dplyr::mutate(part.working = case_when(
  !(v501 %in% c("married","living with partner"))  ~ "no partner",
  v704a == "didn't work last 12 months" ~ "Didn't work last 12 months",
  v704a == "worked last 7 days" ~ "Worked last 7 days",
  v704a == "worked last 12 months" ~ "Worked last 12 months",
  v704a == "don't know" ~ "Don't know"))
```

Skip pattern example: If the respondent does not have a partner, she is not asked about her partner's employment status

03

Clean & Recode Ethiopia DHS Data

activity

Clean and recode Ethiopia DHS data using data cleaning code and univariate output.

04

Consensus on Variables

activity

Develop consensus on a final set of cleaned/recoded variables for EDA

- Look through the histograms and decide on whether any additional changes should be made to the data (e.g., drop or recode variables)
- Input any new variables into the Pathways Workbook and prepare worksheet for EDA step
- Group discussion: Were multiple versions of a variable created through recoding? If so, why?

https://forms.gle/rnHsUT4fkHLyvc1w7

End-of-day survey

05

Office Hours