pathways

Pathways Segmentation Methods Workshop

Day 3 – Exploratory Data Analysis

IDM INSTITUTE FOR DISEASE MODELING

Gates Foundation

Day 3 Outline

- Review from day 2
- What is exploratory data analysis (EDA)
- EDA process
- Activity: Implement the EDA code (in R)
- EDA validation output
- Activity: Variable recoding

Day 3 Outline

- **Regression summary output & decisions**
- **Activity: Decisions about dropping variables**
- **Activity: Debrief activities**

01

Day 2 Review

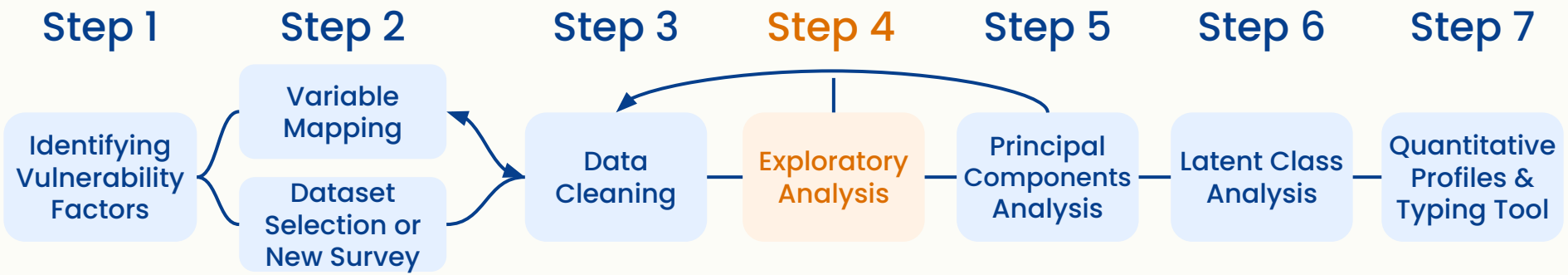
discussion

What did you learn from yesterday?

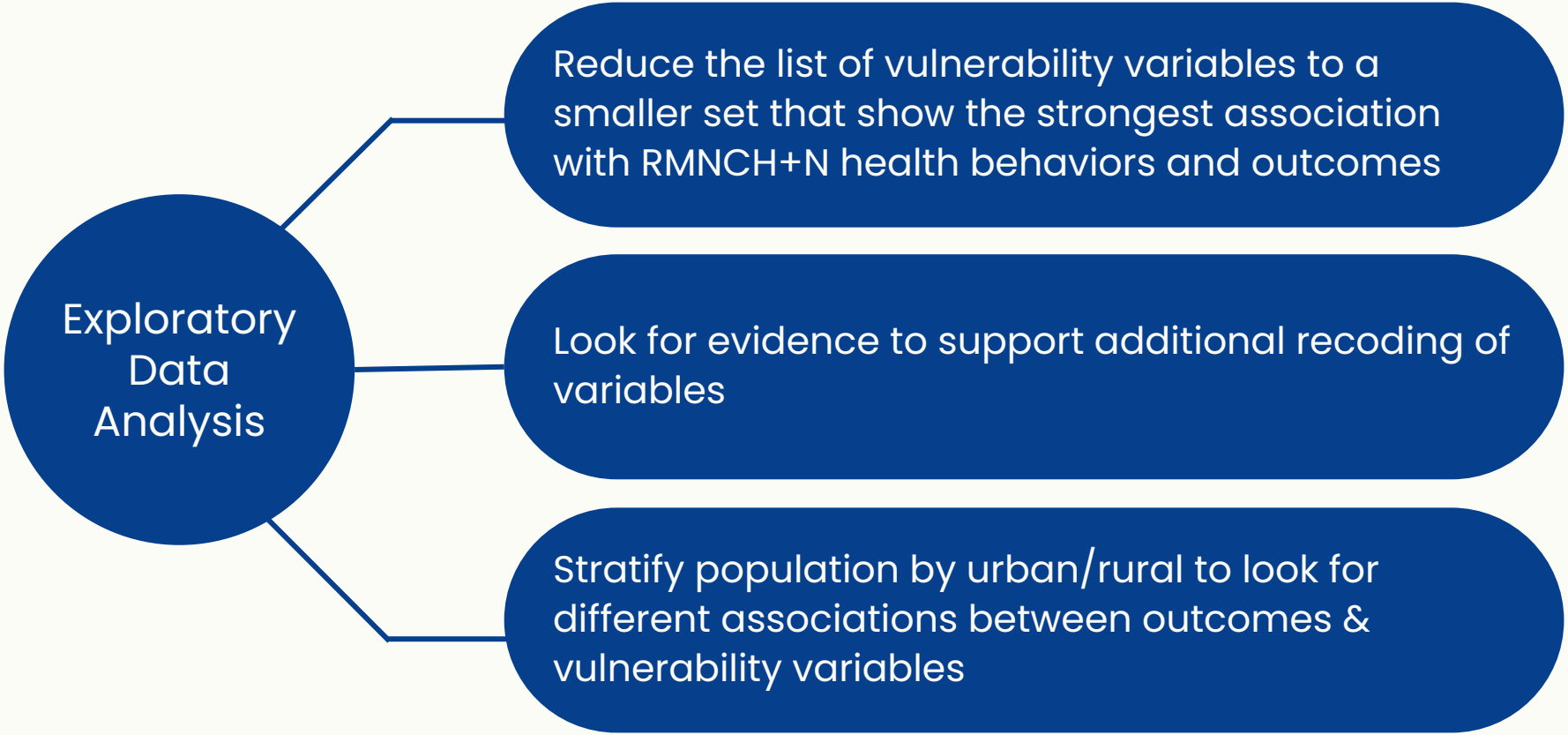
Was there anything confusing from the last two days that you would like to review?

02

What Is Exploratory Data Analysis?


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

What are we doing and why are we doing it?



Why do subgroup analyses for women living in urban vs rural areas?

urban (⅓)

rural (⅔)

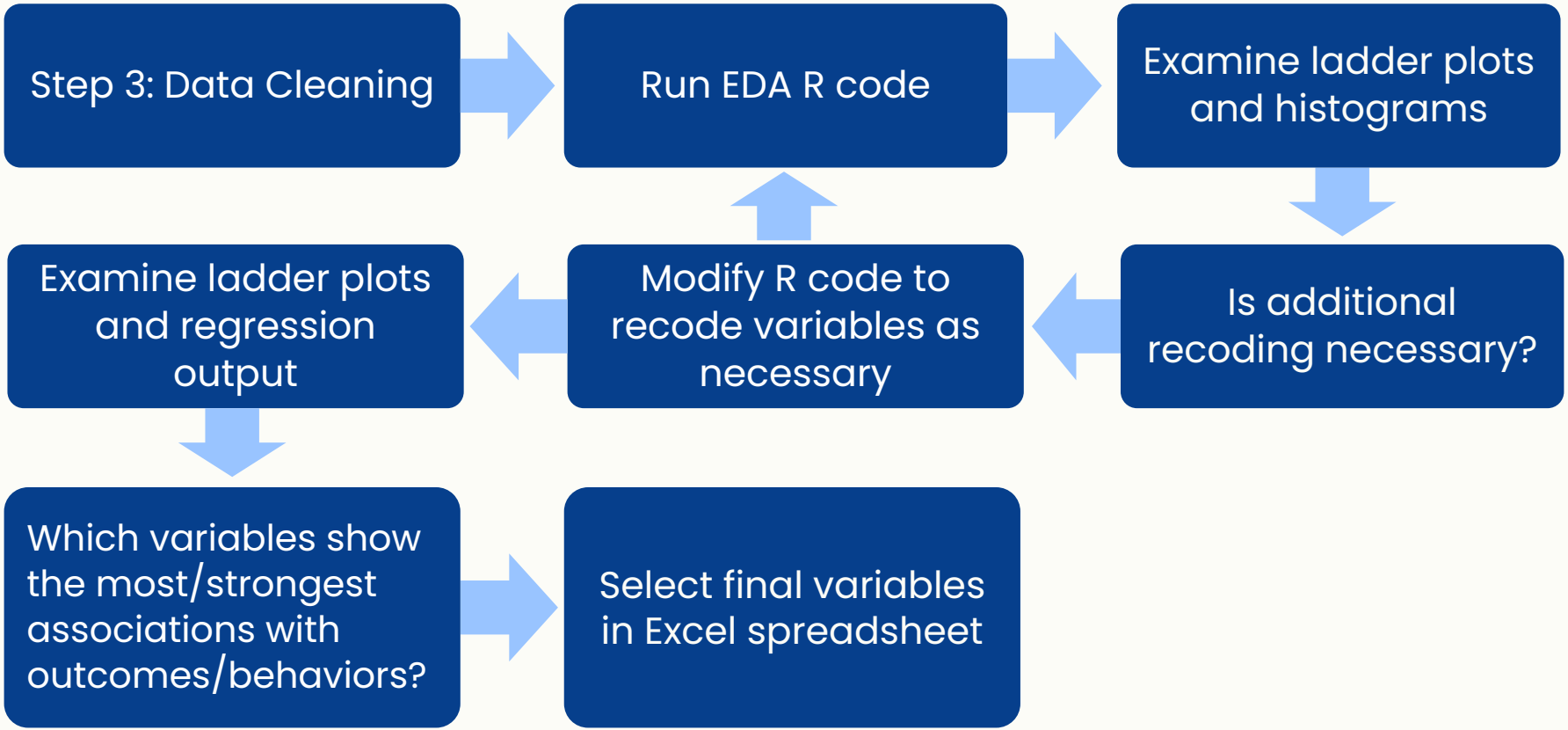
The factors that make women vulnerable to poor health may differ between urban and rural areas!

Pooling data could mask subgroup effects and we could drop vulnerability variables we should keep.

03

EDA Process

EDA Process

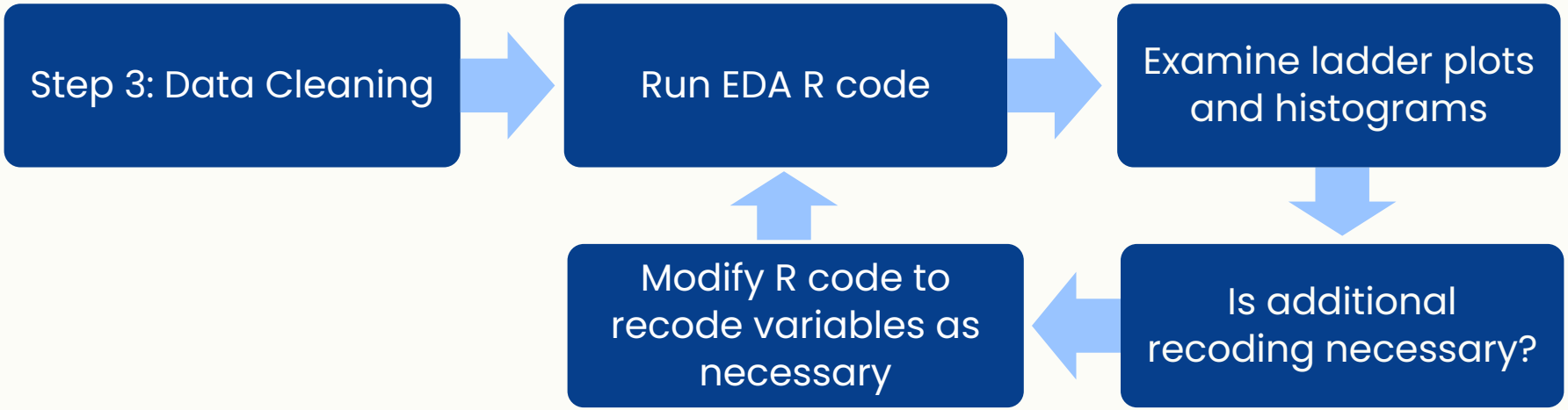

Step 3: Data Cleaning
Run EDA R code
Examine ladder plots and histograms
Is additional recoding necessary?
Modify R code to recode variables as necessary
Examine ladder plots and regression output
Which variables show the most/strongest associations with outcomes/behaviors?
Select final variables in Excel spreadsheet

EDA Process

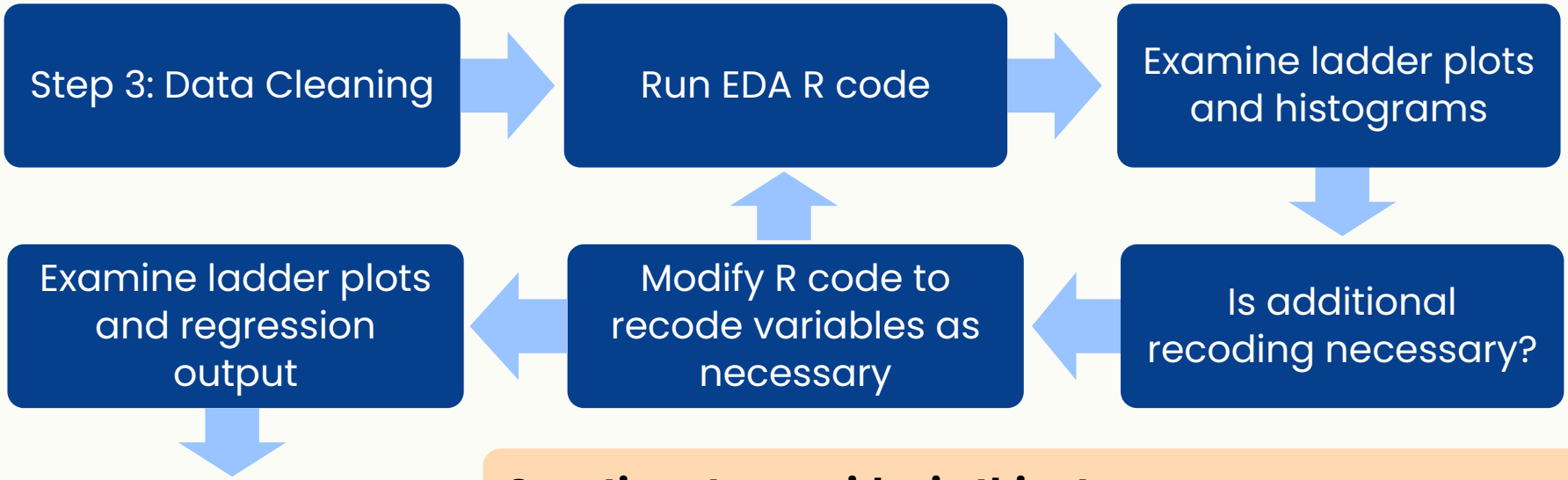
Questions to consider in this step:

- Do histograms/ladder plots show categories with a small sample size or indicate high levels of missing data?
- Do categories of a variable have a similar point estimate in regression models?

EDA Process


Step 3: Data Cleaning
Run EDA R code
Examine ladder plots and histograms
Is additional recoding necessary?
Modify R code to recode variables as necessary

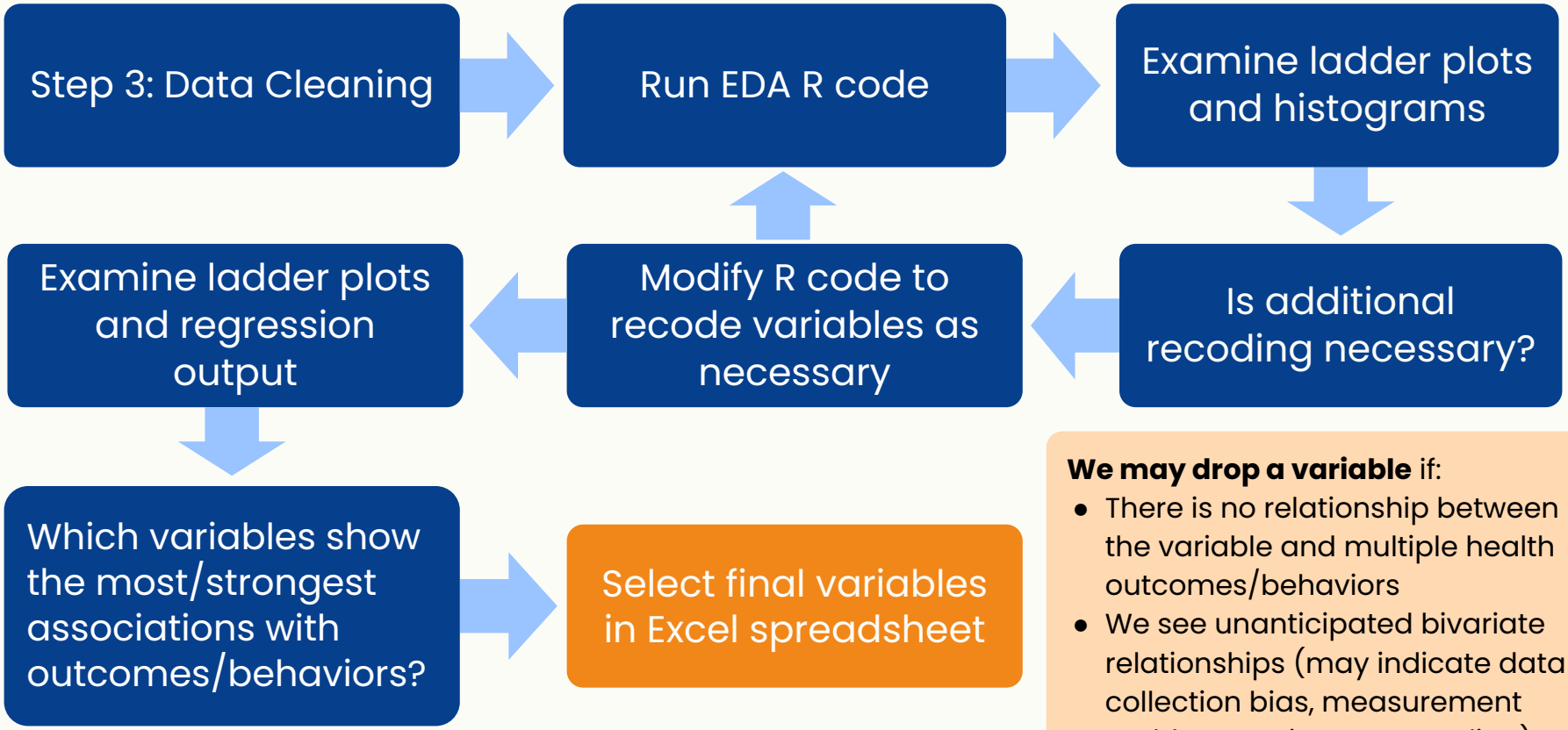
EDA Process



Questions to consider in this step:

- How many significant associations does a vulnerability factor have across all health outcomes/behaviors?
- Does a vulnerability factor have strong associations with one group of outcomes but not another (e.g., only child outcomes)?

EDA Process



We may drop a variable if:

- There is no relationship between the variable and multiple health outcomes/behaviors
- We see unanticipated bivariate relationships (may indicate data collection bias, measurement problems, or incorrect coding)

03

Implement the EDA Code

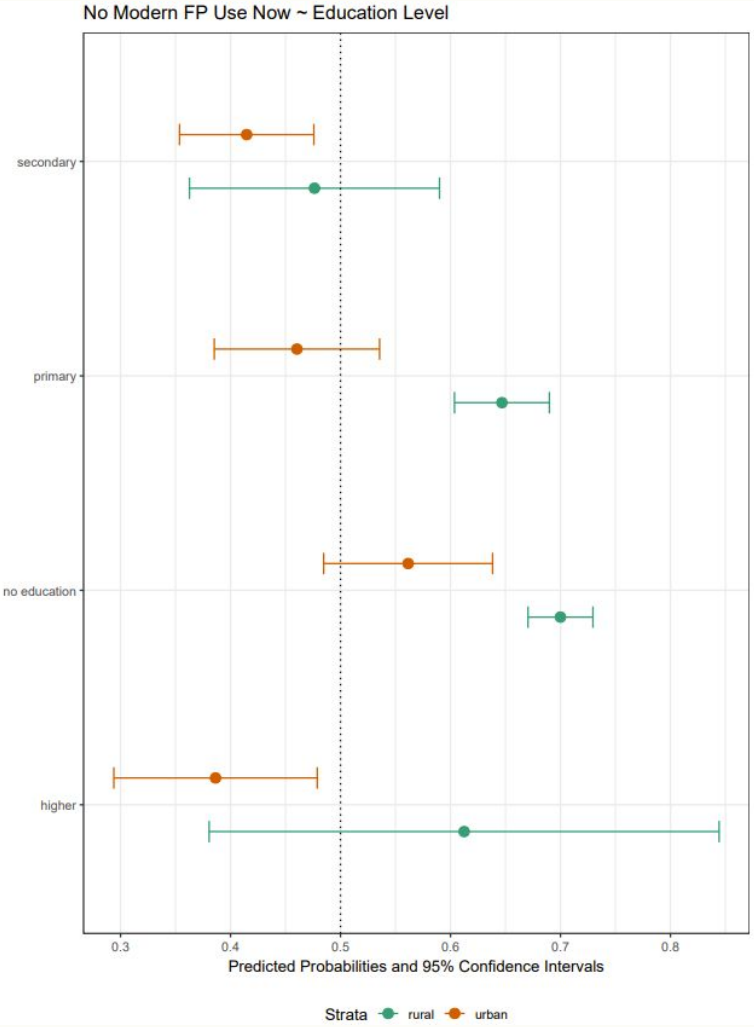
activity

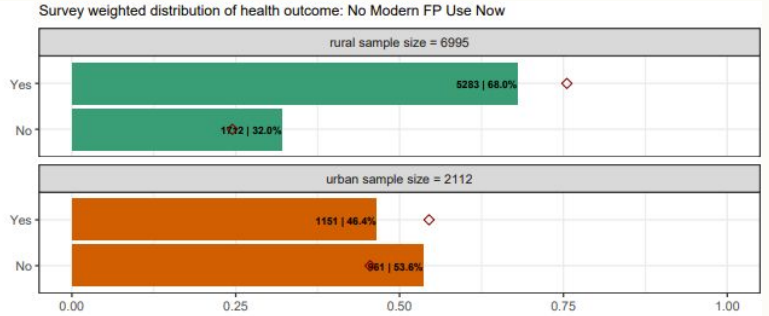
Implement the EDA code (in R) and understand how to use the Pathways Workbook for EDA.

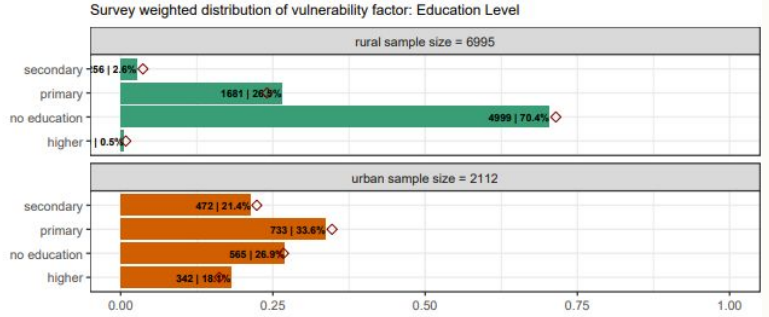
04

EDA Visualization Output

Standard EDA output produced for each outcome~vulnerability variable combination







Binary Outcome variable: nofp.mod.now and categorical Vulnerability Factor: ed.level. Odds ratio displayed.

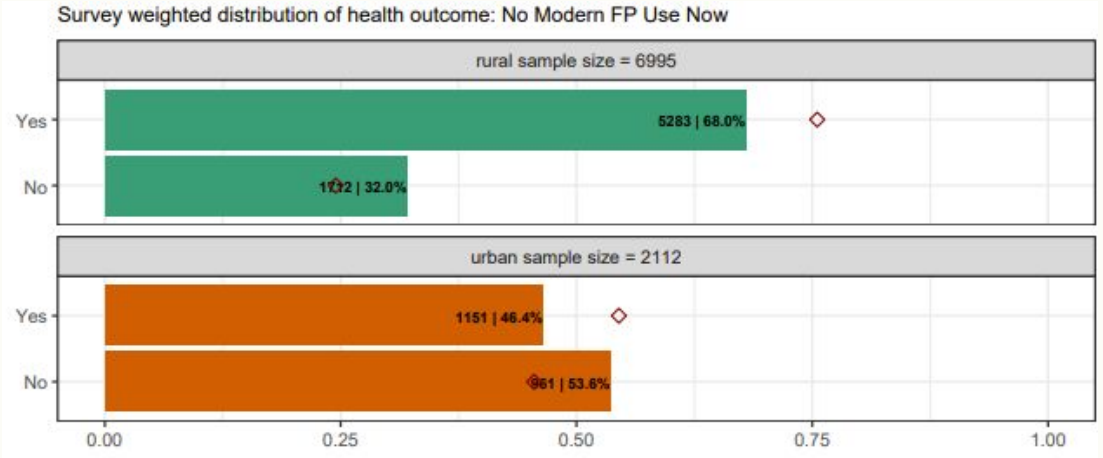
var	rural	urban
secondary	0.39*** [0.24, 0.62]	0.55** [0.36, 0.84]
primary	0.78* [0.65, 0.94]	0.67 [0.43, 1.04]
no education	Ref	Ref
higher	0.68 [0.25, 1.84]	0.49** [0.31, 0.78]

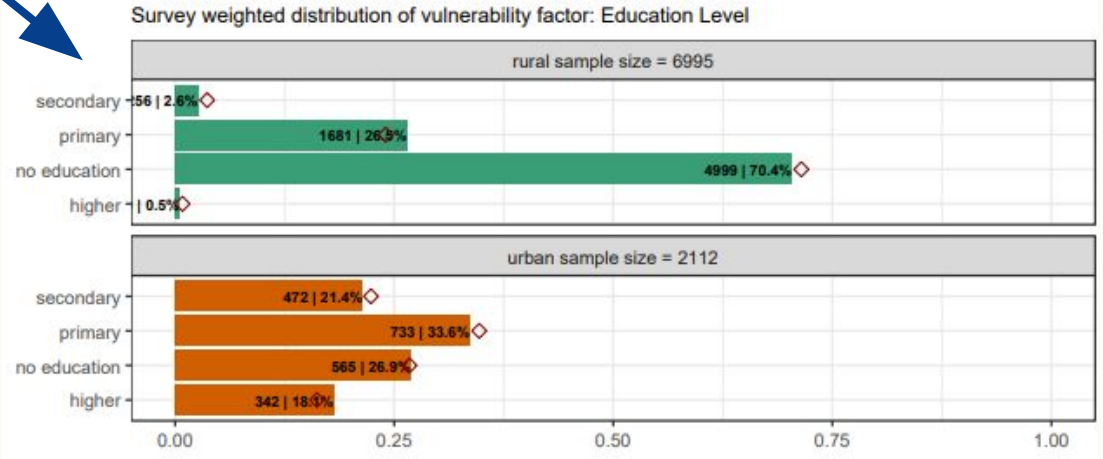
Histograms of the outcome and vulnerability variable

- Are the variables coded correctly?
- Do histograms show categories with a small sample size?
- Do histograms indicate there may be a lot of missing data?

Do we need to do additional recoding of this variable?

Is there evidence to support dropping this variable?



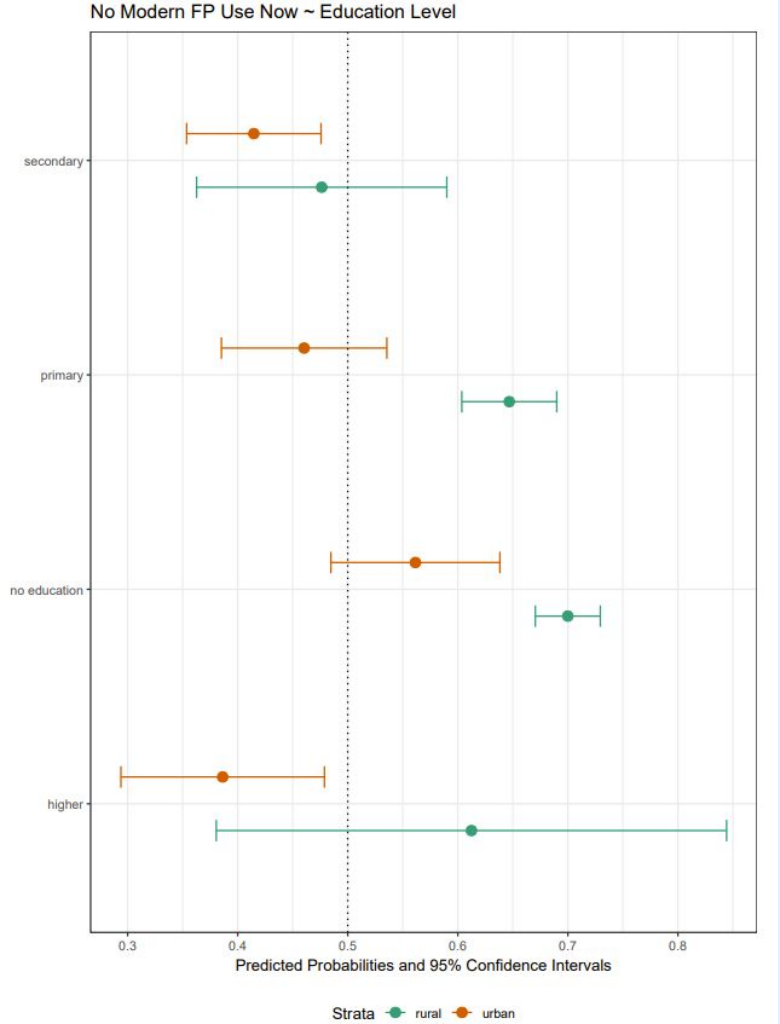


Ladder plot show predicted probabilities and 95% CIs from bivariate regression

- Do ladder plots show categories with a small sample size or indicate high levels of missing data?
- Do categories of a variable have a similar point estimate in regression models?

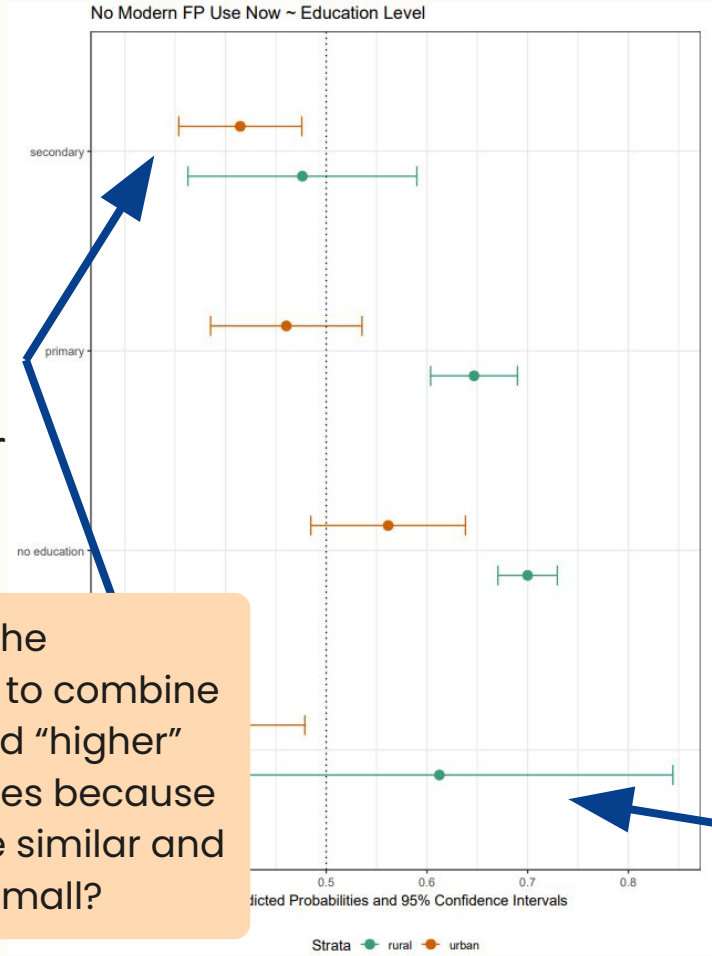
Should we recode the variable to combine multiple categories because point estimates are similar?

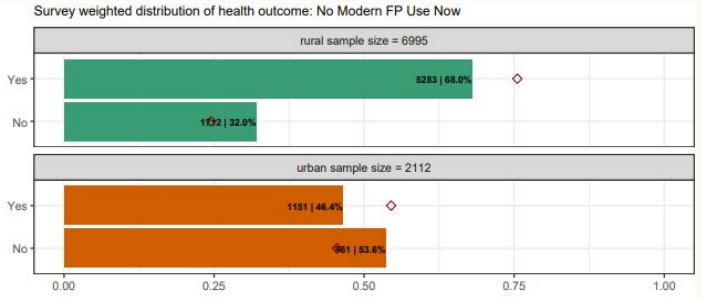
Is there evidence to support dropping this variable?

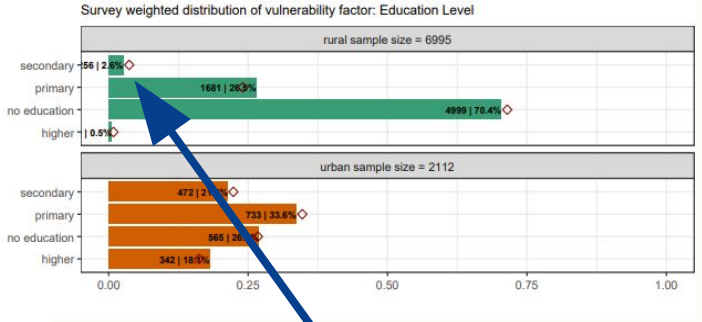


Example decisions made with EDA output

Point estimates for the “secondary” and “higher” education categories are similar







Should we recode the education variable to combine the “secondary” and “higher” education categories because point estimates are similar and the sample size is small?

In the rural sample, we have a very small sample size the “higher” education category

activity

Review **histograms** and **ladder plots** from EDA visualization output to determine if any additional variable recoding is required.

Loop back to data cleaning R code and **do additional recoding;** include new variables in the Pathways Workbook.

05

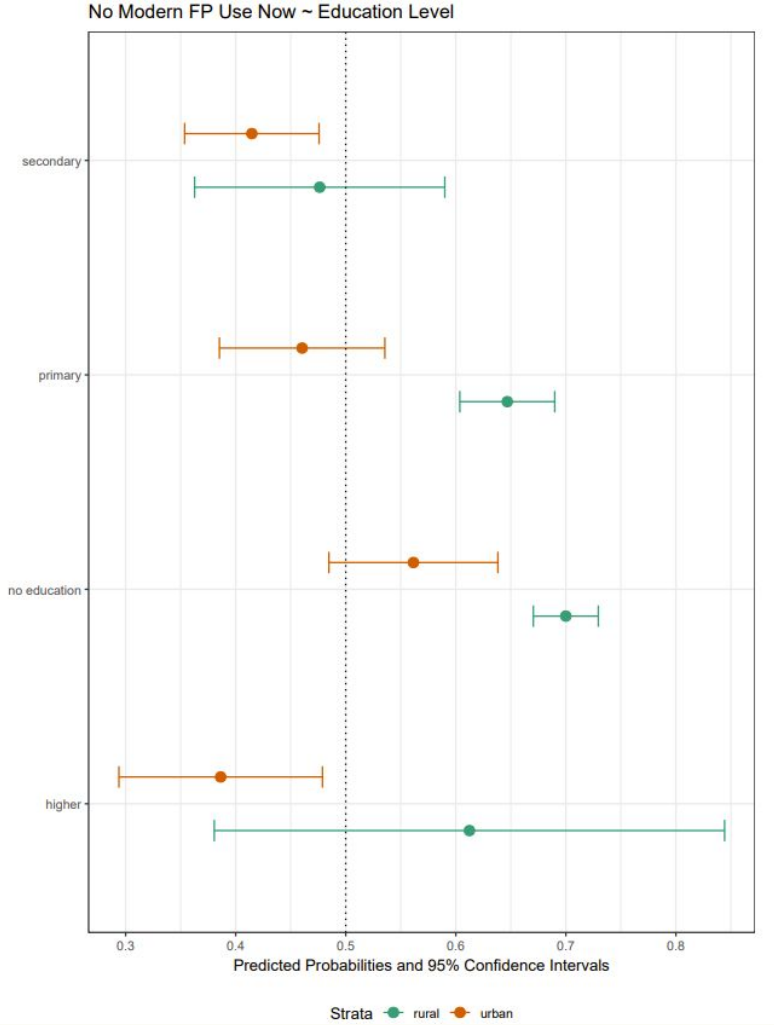
Regression Summary Output & Decisions

A regression table shows the Odds Ratios and 95% CIs from bivariate regression

- How do we interpret the predicted probabilities in the ladder plot?
- What does this output tell us about the association between the outcome and vulnerability?

Binary Outcome variable: nofp.mod.now and categorical Vulnerability Factor: ed.level. Odds ratio displayed.

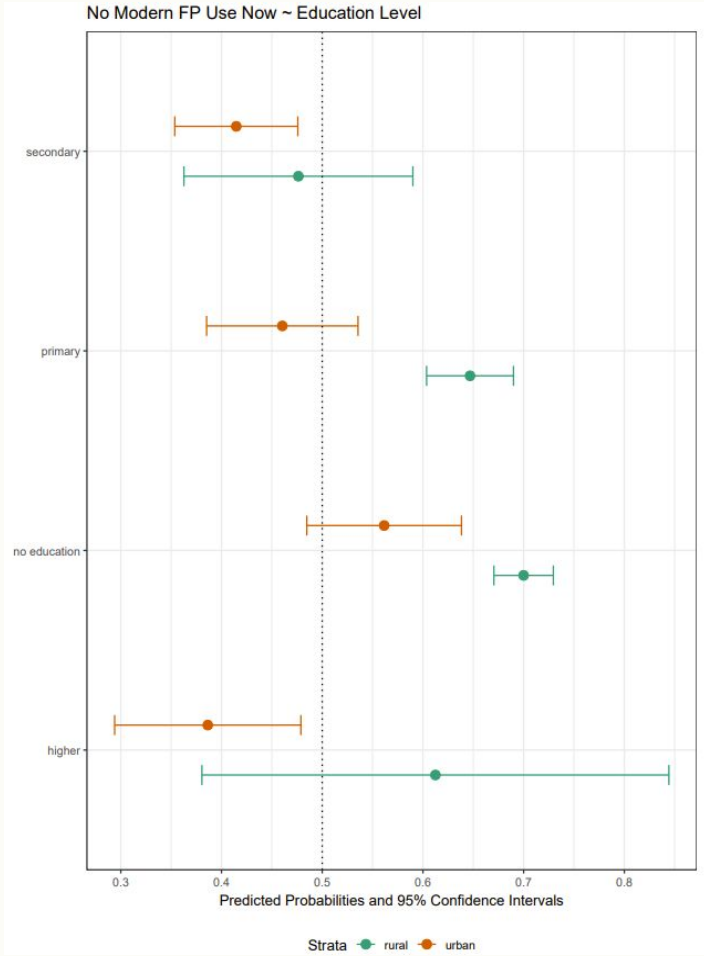
var	rural	urban
secondary	0.39*** [0.24, 0.62]	0.55** [0.36, 0.84]
primary	0.78* [0.65, 0.94]	0.67 [0.43, 1.04]
no education	Ref	Ref
higher	0.68 [0.25, 1.84]	0.49** [0.31, 0.78]

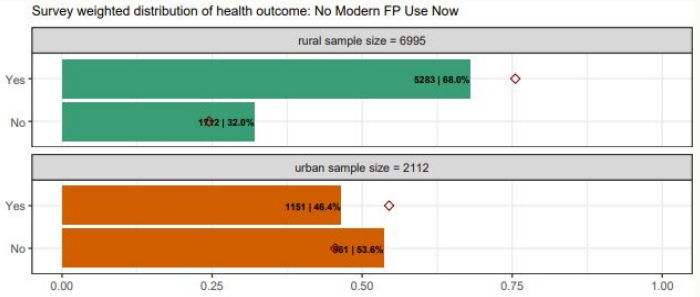


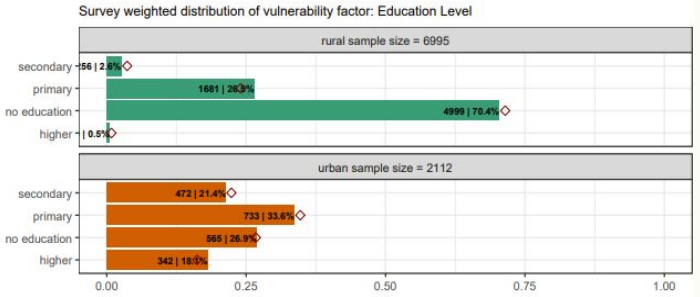
Example decision made with EDA regression output

Is there a significant association between education and modern FP use?

Is this association significant for both urban and rural populations?







Binary Outcome variable: nofp.mod.now and categorical Vulnerability Factor: ed.level. Odds ratio displayed.

var	rural	urban
secondary	0.39*** [0.24, 0.62]	0.55** [0.36, 0.84]
primary	0.78* [0.65, 0.94]	0.67 [0.43, 1.04]
no education	Ref	Ref
higher	0.68 [0.25, 1.84]	0.49** [0.31, 0.78]

We must look across multiple outcomes before we can make a final decision about keeping/dropping a vulnerability variable at this stage!

- How many significant associations does a vulnerability factor have across all health outcomes/behaviors?
- Does a vulnerability factor have strong associations with one group of outcomes but not another (e.g., only child outcomes)?

activity

Review all regression summary output and make final decisions about dropping variables.

Record decisions in the Pathways Workbook.

activity

Debrief today’s activities.

https://forms.gle/UpDrBtNa5CJoNGS49

End-of-day survey

pathways

Pathways Segmentation Methods Workshop

Day 4 – Variables

IDM

Gates Foundation

Day 4 Outline

- Review from day 3
- Recoding & dropping variables with EDA output
- Activity: Variable coding & finalization
- Activity: Variable presentations
- Activity: Debrief

01

Day 3 Review

discussion

What did you learn from yesterday?

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

02

Recoding & Dropping Variables Using EDA Output

Elisabeth & Jeremy to add slides as needed

03

Variable Coding & Finalization

activity

Finalize variable coding and develop final set of variables for PCA.

Record decisions in Pathways Workbook.

04

Variables Presentations

activity

Present the final set of variables selected for the PCA, including how variables were recoded.

05

Debrief

discussion

Do we have an agreed upon set of health outcomes and vulnerability factors?

https://forms.gle/3gkNASMzp69xRWVx7

End-of-workshop survey