pathways

Pathways Segmentation Methods Training

Virtual Session 1 – Principal Components Analysis

IDM

Gates Foundation

Session 1 Outline

- Principal Components Analysis (PCA)
- PCA process
- Reviewing output: what can the PCs tell us?

01

Workshop Review

discussion

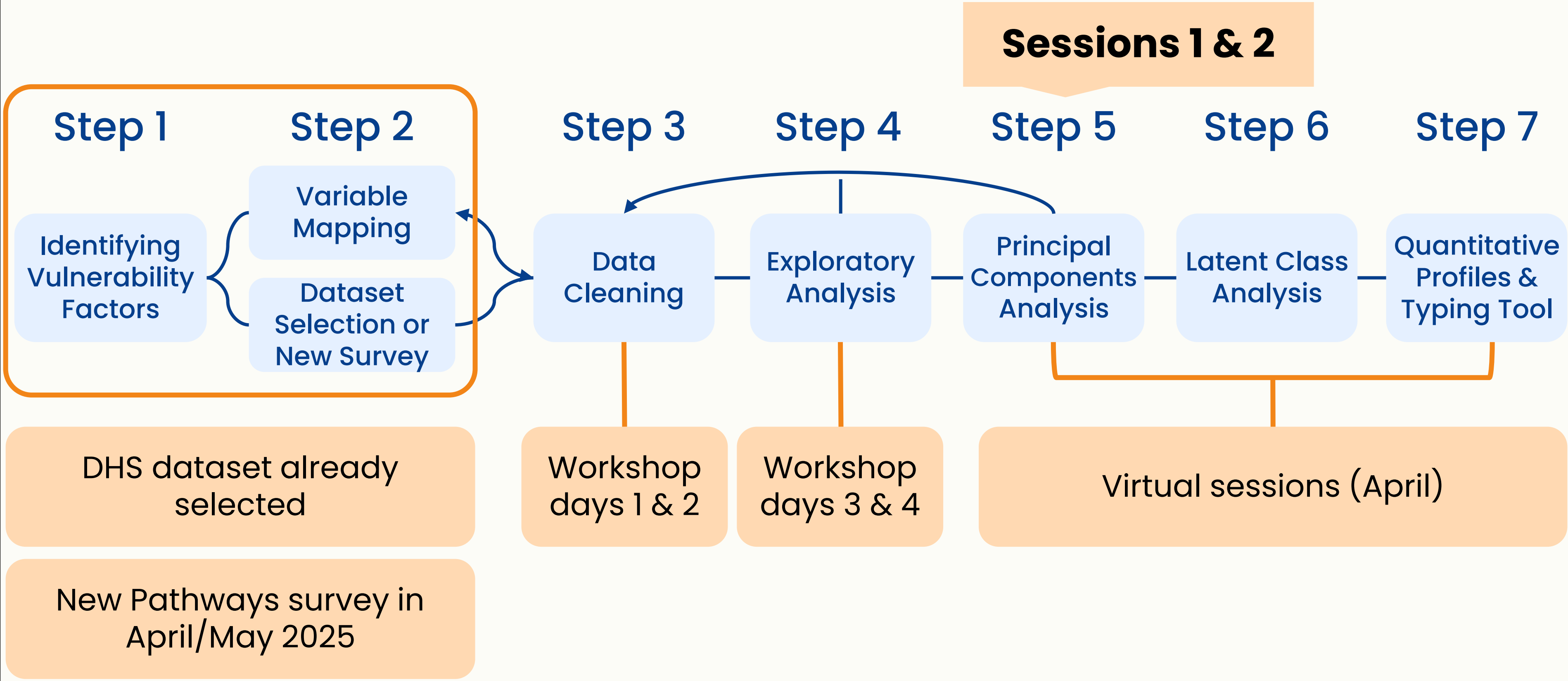
Have you practiced the first two steps of the data segmentation process?

Do you have any questions?

02

Pathways Segmentation Method

Timeline for segmentation training



Principal Components Analysis

PURPOSE

To reduce the number of variables representing each vulnerability domain by removing some variables that are strongly correlated (i.e., measure the same thing).

INPUTS

Subset of variables identified in exploratory data analysis step.

DECISIONS

- What variables to include in LCA
- A list of highly correlated variables that were dropped, that could be added back if LCA performance is not good

Principal Components Analysis

CONSIDERATIONS

Correlation between variables by domain.

DECISION SUPPORT TOOLS

- Scree plots
- Bi-plots
- Variable composition plots

OUTPUTS

Excel spreadsheet with indicator of the reduced set of variables to pass to the LCA.

03

PCA

Why do we need PCA?

As datasets grow in complexity, **they often contain a large number of correlated variables**

For example, economic empowerment can be measured by:

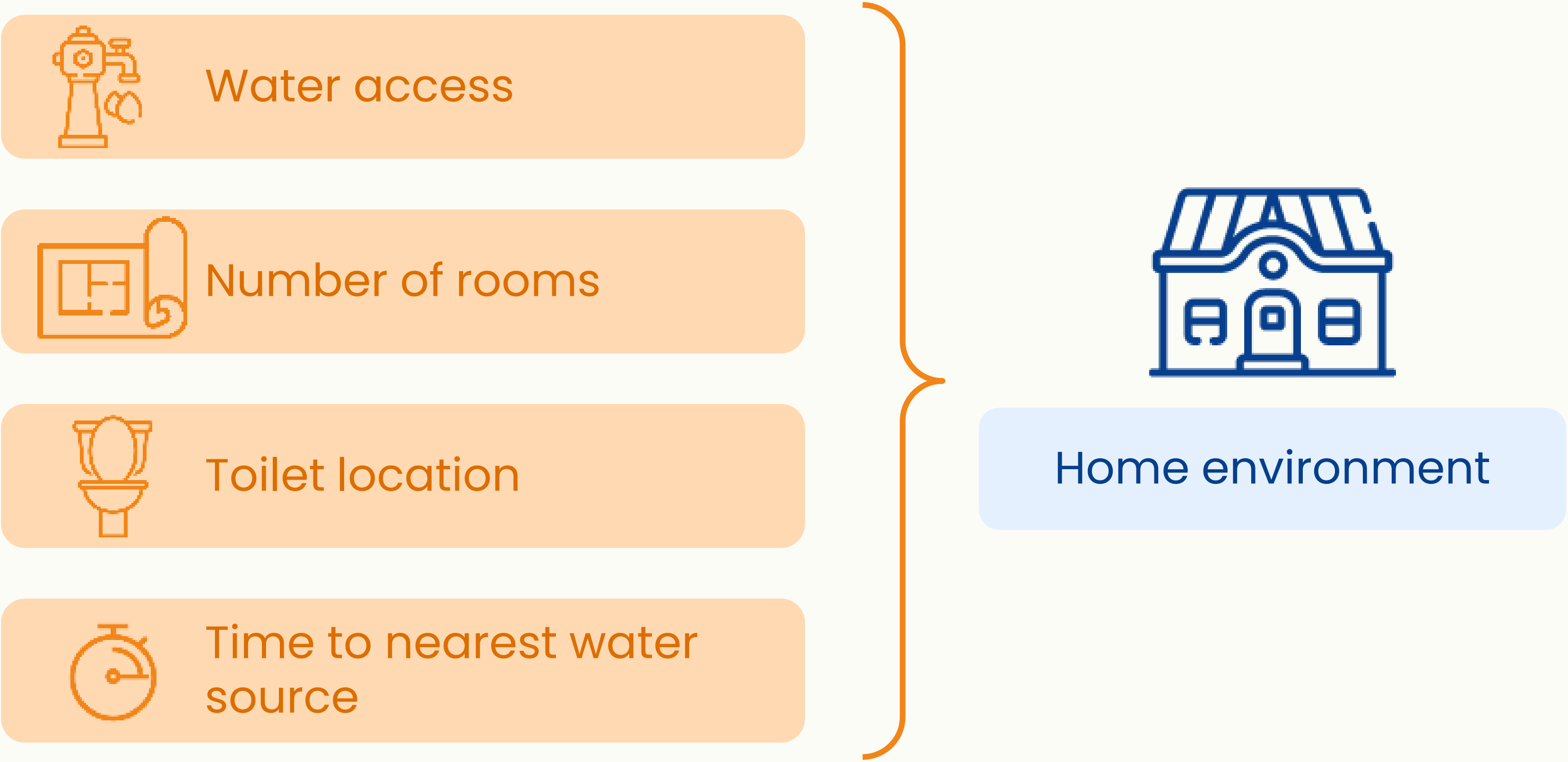
- A woman's employment status
- What type of work she does
- Her education level
- Whether she has a bank account
- If she has control over her own income

All these indicators provide some information but since they are correlated, **a select few are sufficient** to describe economic empowerment

Having many correlated variables creates many problems in modeling:

- **Redundancy**: Highly correlated variables provide the same information and can increase the complexity of the model without adding value
- **Interpretability**: A large number of variables make it challenging to interpret the final model
- **Generalizability**: Models with highly correlated variable often "overfit" the data and are difficult to generalize to new datasets
- **Computational Efficiency**: Models with many variables are computationally intensive and take longer to run

PCA addresses these issues by **reducing the dimensionality of the dataset – e.g., dropping correlated variables –** while retaining the most important patterns in the data


Water access
Number of rooms
Toilet location
Time to nearest water source
Home environment

Water access

Number of rooms

Toilet location

Time to nearest water source

These variables are often **correlated** and/or **redundant** since they are meant to measure the same thing

Home environment

PCA:

a statistical technique used to reduce the number of variables in a dataset while preserving as much variability (information) as possible

principal components (PCs):

a weighted linear sum of the original set of variables

PCs

PC example

PC = 0.5*toilet_location + 0.4*water_source+ 0.2*rooms_in_house

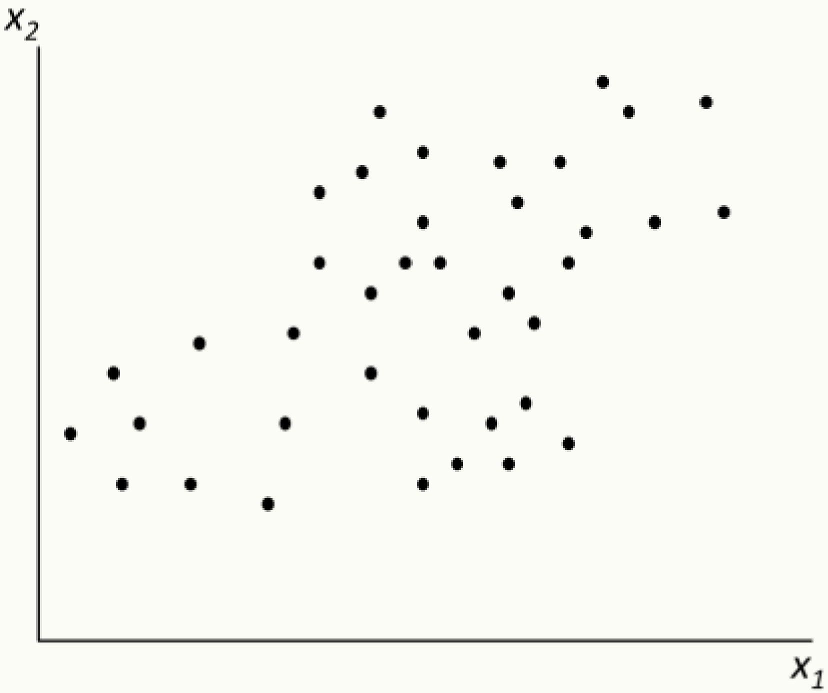
Each principal component is a linear combination of the original variables.

The **loading** is the coefficient (or weight) of a variable in that linear combination.

- The loadings for this PC are: 0.5, 0.4, and 0.2
  - These tell you **how strongly each variable influences that PC**

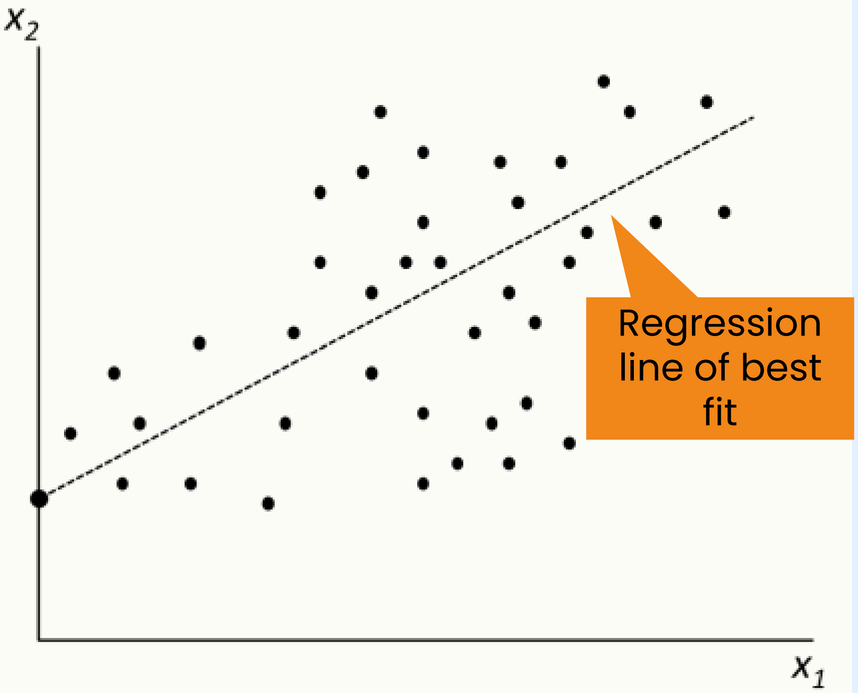
How do we determine these loadings? And how many PCs do we need?

How does PCA work?



An iterative process, where each step aims to explain more and more variation in the data

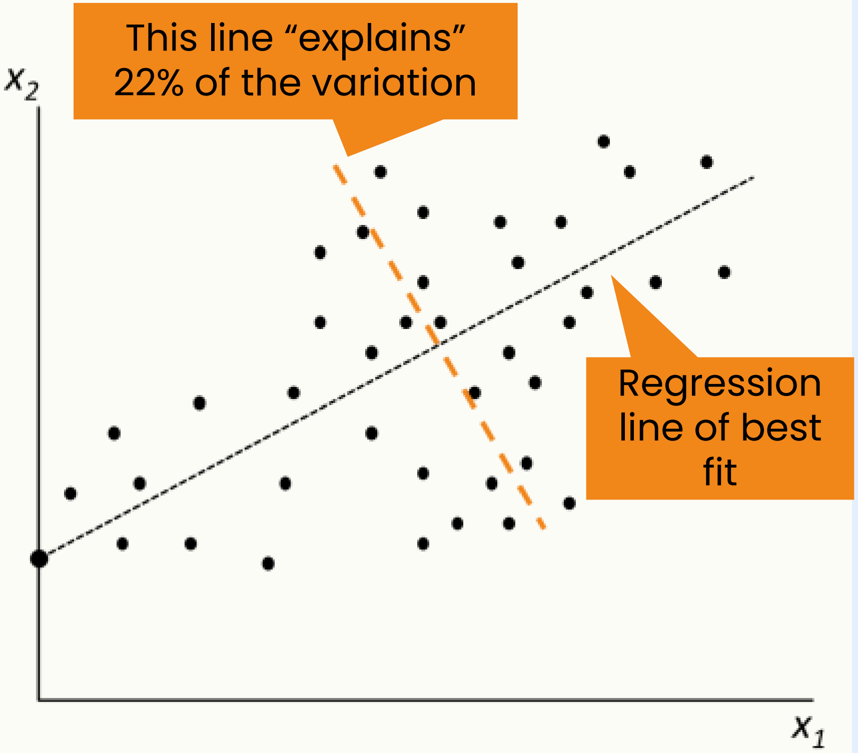
How does PCA work?



1. Find best fitting linear combination of variables that best explains the patterns in the data (PC1).

"Explains" 50% of the variation In the data

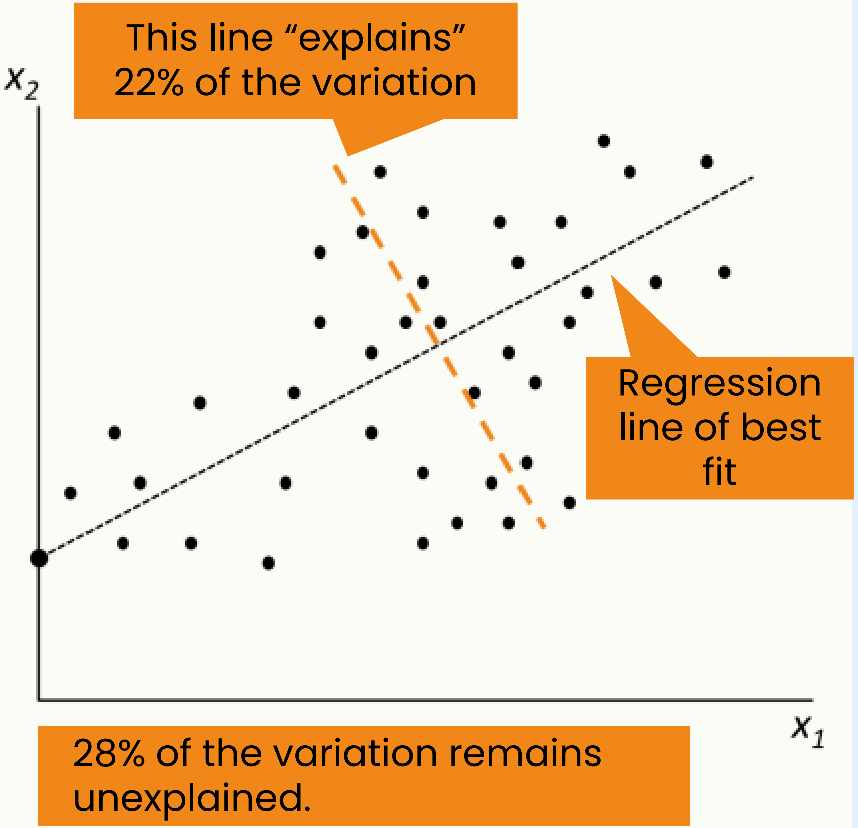
How does PCA work?



1. Find best fitting linear combination of variables that best explains the patterns in the data (PC1).

2. Fit a new line, orthogonal to PC1, that best explains the remaining variation in the data (PC2).

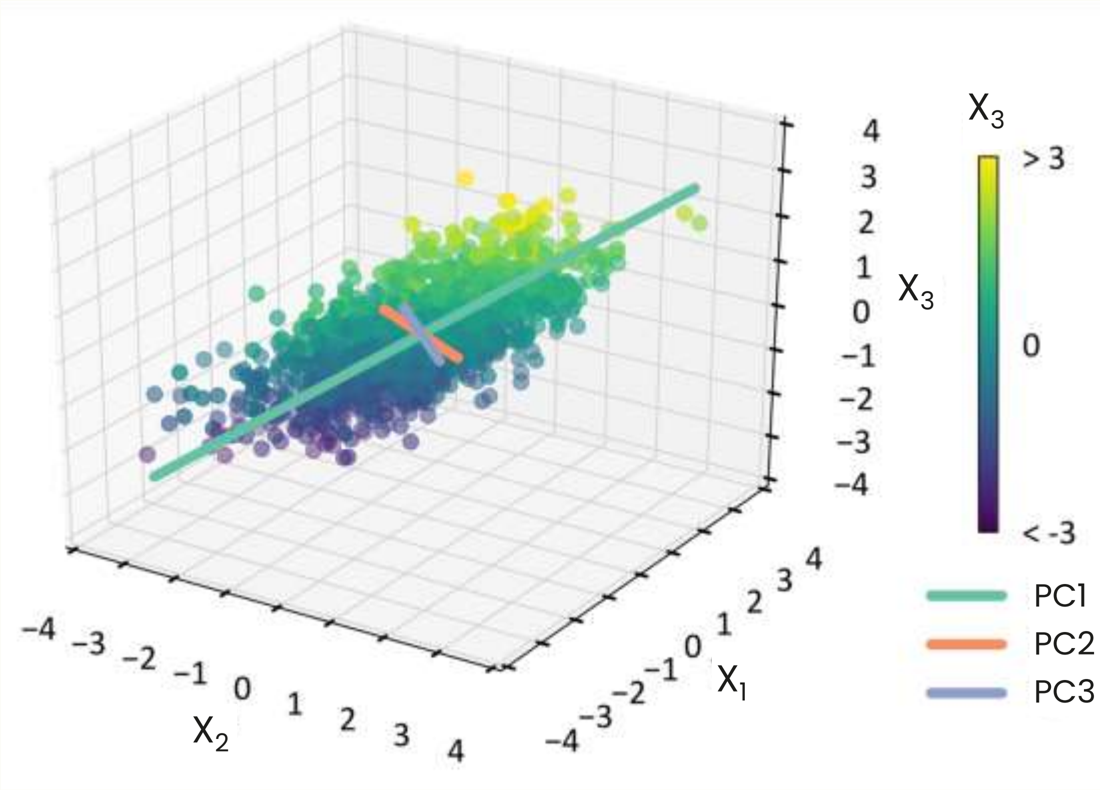
How does PCA work?



1. Find best fitting linear combination of variables that best explains the patterns in the data (PC1).
2. Fit a new line, orthogonal to PC1, that best explains the remaining variation in the data (PC2).
3. Continue this process until you have enough PCs to explain a sufficient amount of variation in your data.

With a many variables, we typically need more than 2 PCs to explain the variation in the data

But, *n*-dimensional space is very difficult to work in, so we typically compare two PCs at a time

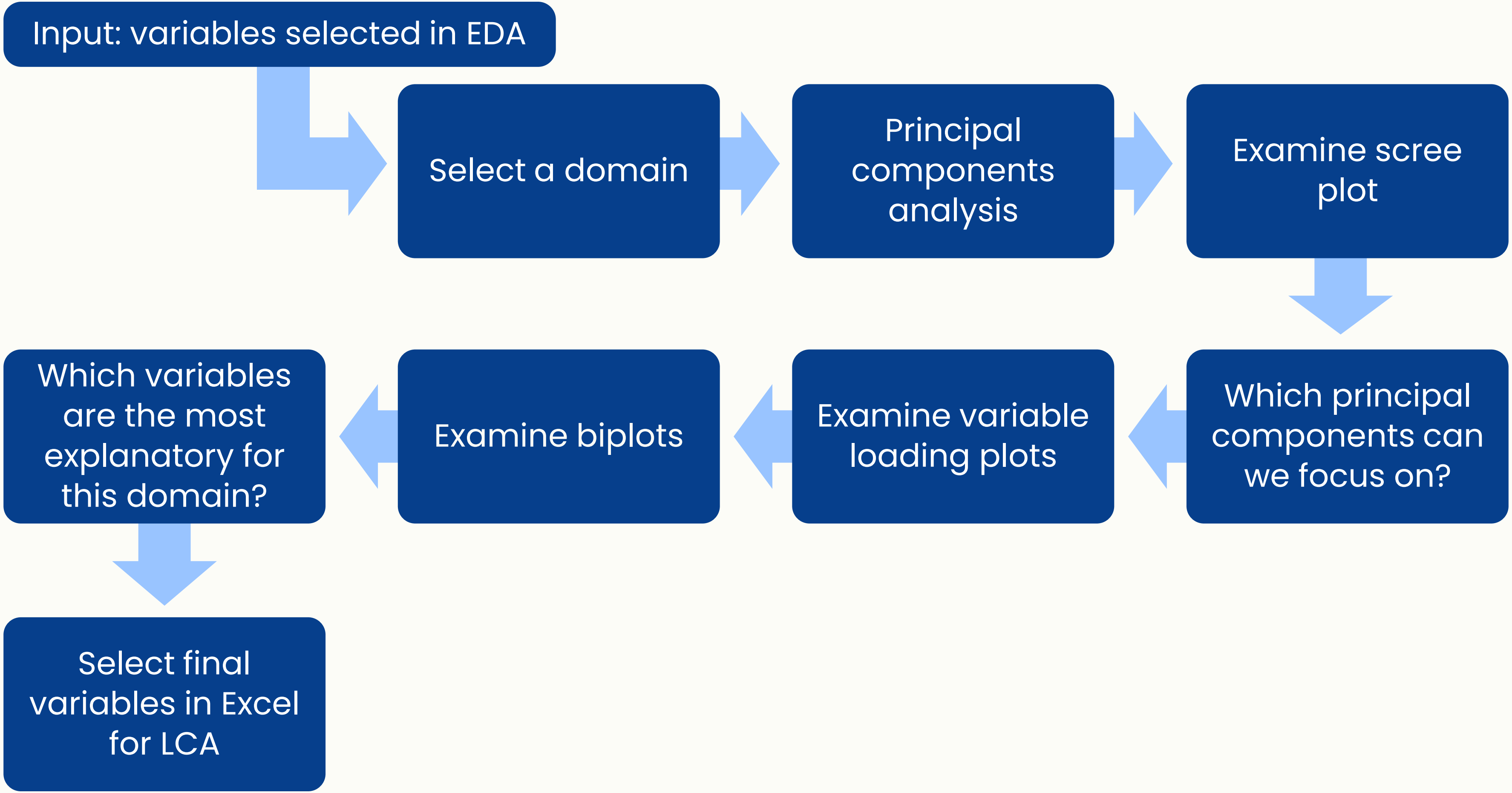


PC1 explains the most variation in the data

PC2 and PC3 explain a much smaller amount of the variation

04

PCA Process


Input: variables selected in EDA
Select a domain
Principal components analysis
Examine scree plot
Which principal components can we focus on?
Examine variable loading plots
Examine biplots
Which variables are the most explanatory for this domain?
Select final variables in Excel for LCA

Input: variables selected in EDA

Select a domain

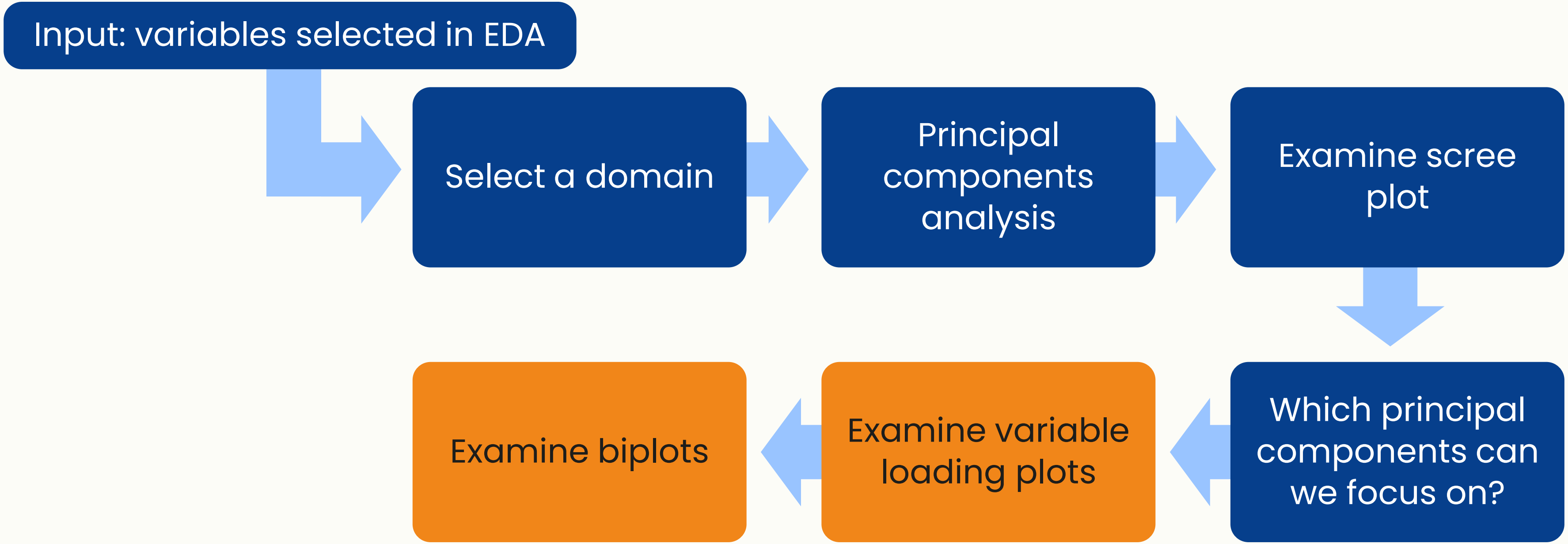
Principal components analysis

Examine scree plot

Which principal components can we focus on?

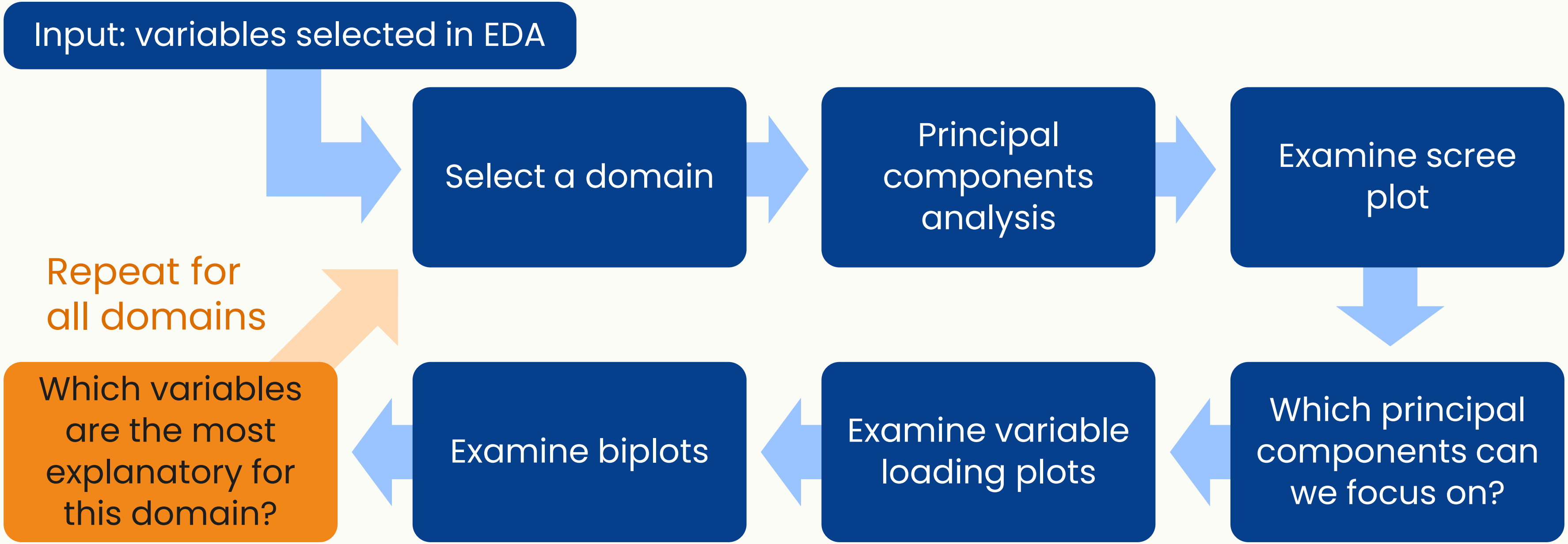
Questions to consider in this step:

- How much variation is explained by each PC?
- How many components are needed to explain ~60% of the variation in the data?



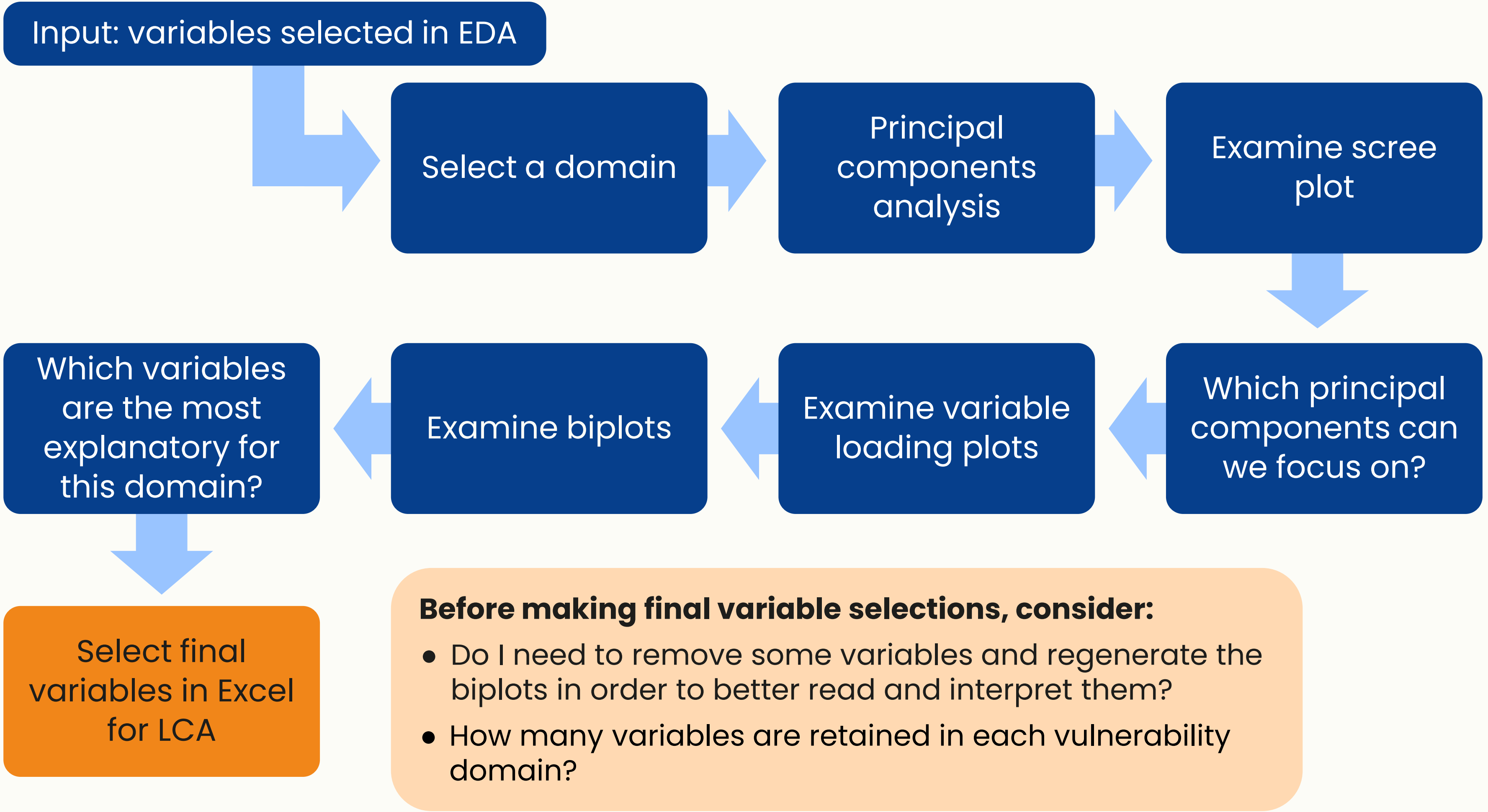
Questions to consider in this step:

- Which variables contribute the most to the PCs/have the highest loadings?
- Which variables appear to be strongly positively/negatively correlated on the biplots? Which are not correlated with each other?
- For variables with more than 2 response categories, how are all of them correlated with other variables in a domain?



Questions to consider in this step:

- When choosing between two or more correlated variables, is there a variable that is easier to interpret or intuitively makes more sense in the context or setting?
- When choosing between two or more correlated variables, is there a variable that is more strongly associated with health outcomes?


Input: variables selected in EDA
Select a domain
Principal components analysis
Examine scree plot
Which principal components can we focus on?
Examine variable loading plots
Examine biplots
Which variables are the most explanatory for this domain?
Select final variables in Excel for LCA
Before making final variable selections, consider:
• Do I need to remove some variables and regenerate the biplots in order to better read and interpret them?
• How many variables are retained in each vulnerability domain?

05

Reviewing Output: What Can the PCs Tell Us?

information from PCA useful for reducing number of variables

How many PCs sufficiently explain the patterns in our data?

information from PCA useful for reducing number of variables

How many PCs sufficiently explain the patterns in our data?

Focus on the first few PCs that explain "most" of the variation in our data.

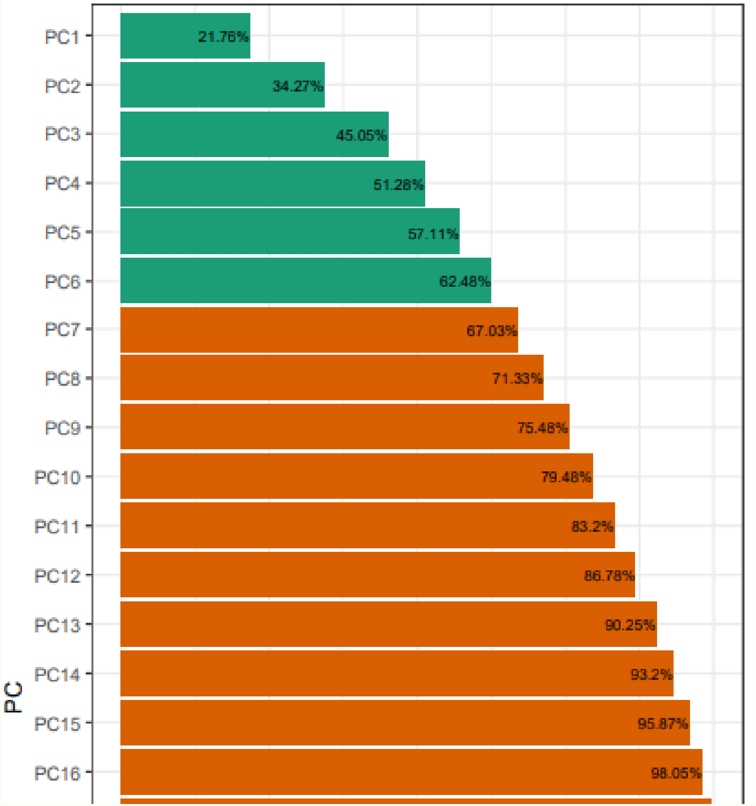
A rule of thumb is to look at the PCs that cumulatively account for ~60% of the variation.

scree plots:

a visual tool used to help determine how many principal components to retain

Scree Plots

Cumulative proportion of variance explained

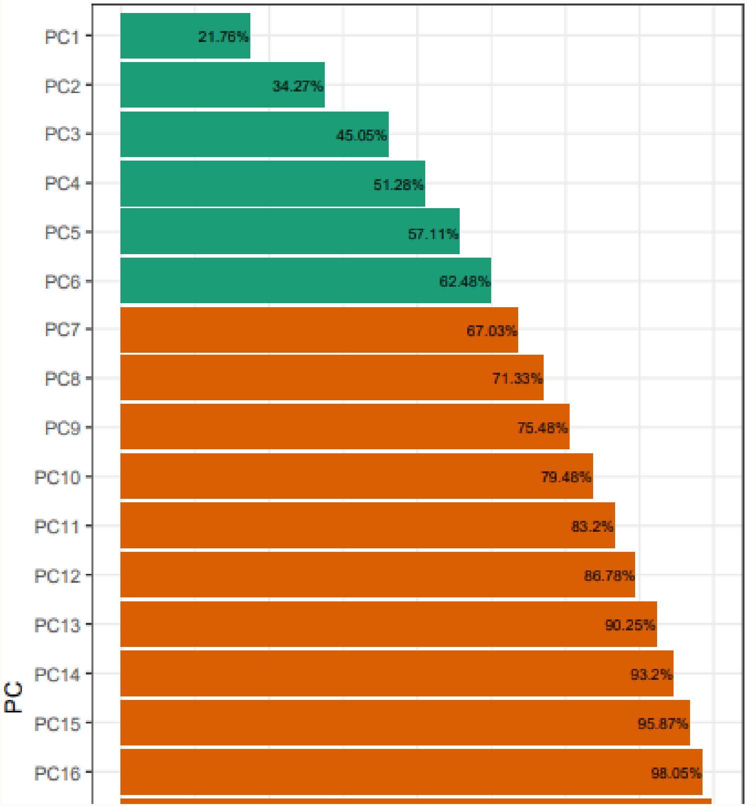


N PC=28

22% of variation in explained by PC1 alone

Scree Plots

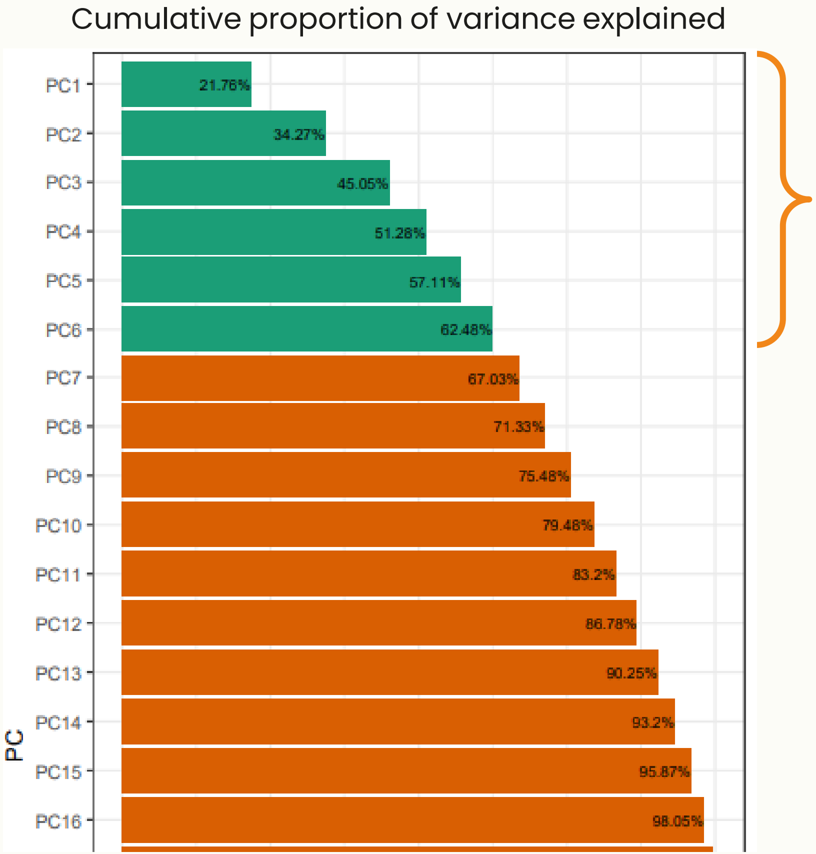
Cumulative proportion of variance explained



22% of variation in explained by PC1 alone

Use as many PCs as needed to capture ~60% of the variation in the data

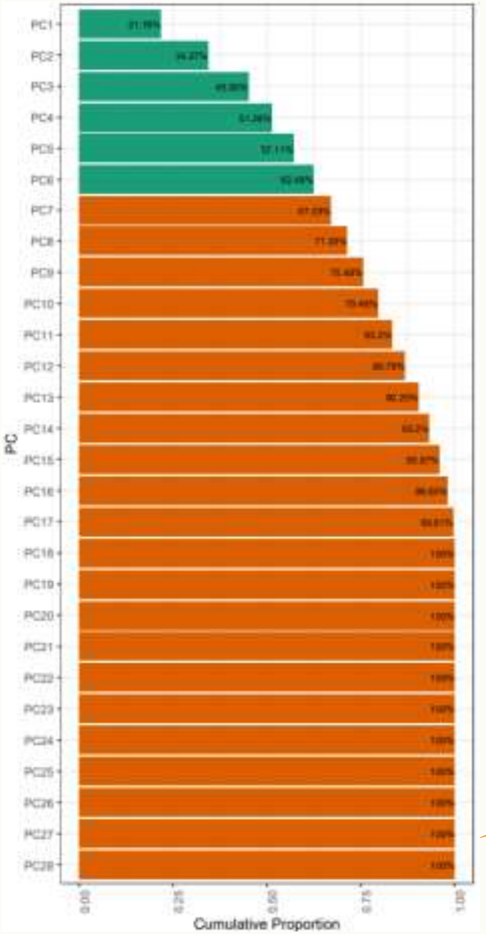
Scree Plots



Between 5 and 6 PCs necessary to explain about 60% of the variation in the data

These 6 PCs explain enough variation that we can just focus on them to identify the most important variables

Scree Plots



If we used all the PCs, we can explain all the variation in our data, but this is equivalent to working with all our original variables (no dimension reduction).

information from PCA useful for reducing number of variables

How many PCs sufficiently explain the patterns in our data?

How do each of the original variables contribute to the PC?

The variables that are more heavily weighted in the PCs are more explanatory of patterns in the data.

Variable Loadings of the PCs

$$PC_1 = 0.44\underbrace{working}_{} + 0.40 wealth\_index + 0.47 somevar + 0.03 somevar2$$

Large loading (0.44) contributes substantially to PC1

This variable is very helpful in explaining the variation in the data.

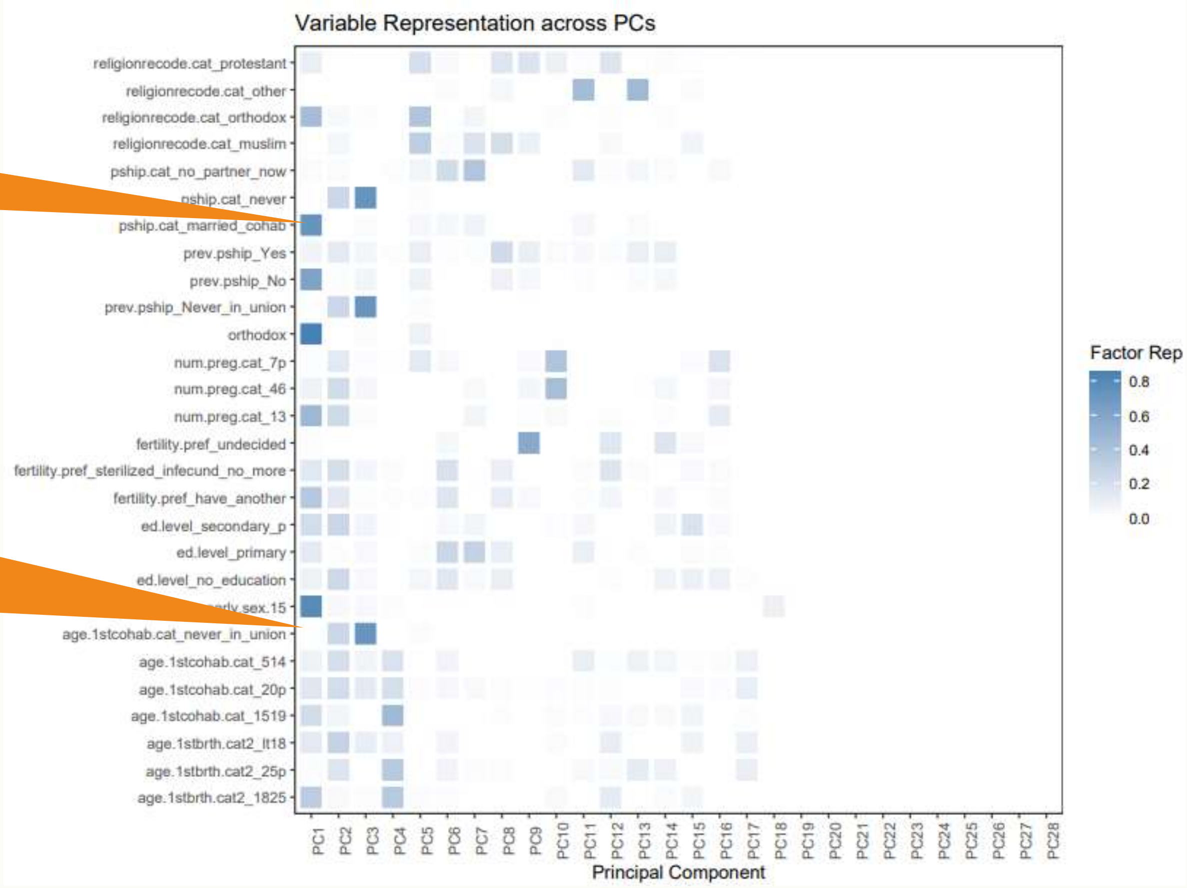
Small loading (0.03) not as helpful in explaining the variation in the data in PC1

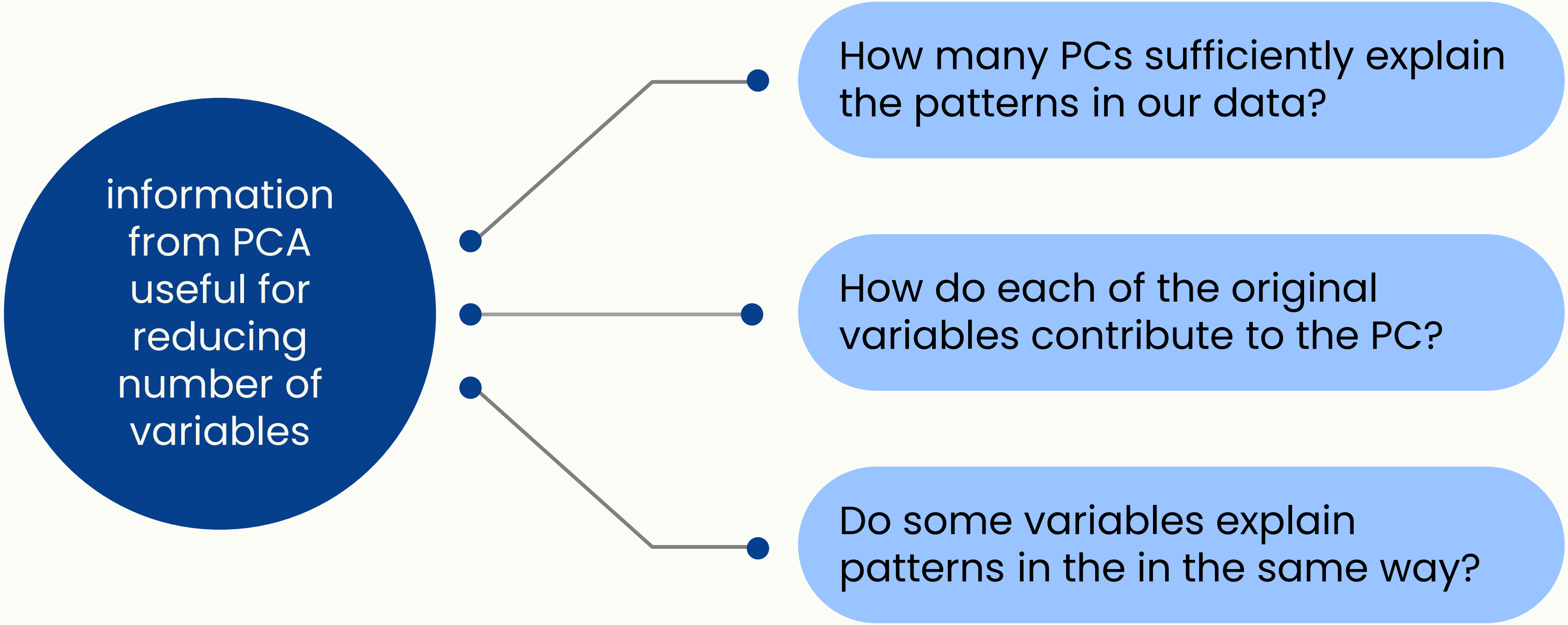
But it could have a large loading in PC2 or PC3, which also capture patterns in our data.

Factor Representation Plot

Contributes quite a bit to PC1

Contributes very little to PC1
BUT contributes a lot to PC3





Variables that are collated within and across PCs are likely redundant, so we can choose one.

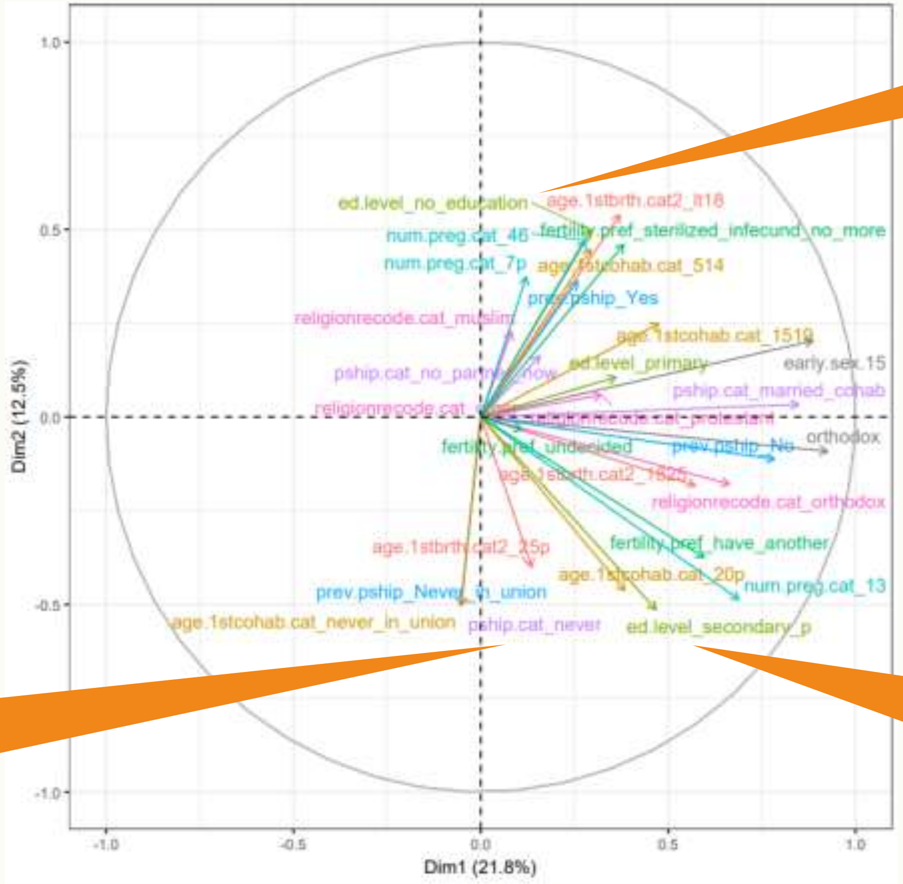
bioplots:

show the correlation between variables across any two PCs, and tell us which variables explain variation in the same way

Bioplots: Direction

Positively correlated variables	point in same direction	explain the same variation in the same way
Negative correlated variables	point in opposite direction	explain the same variation in opposite ways
Uncorrelated variables	perpendicular	explain different patterns in the data (not redundant)

Example



No education and age at first birth < 18 **positively correlated**

Religion muslim is **negatively correlated** with having never been in a partnership

Secondary and higher education **uncorrelated** with religion = muslim and is **slightly correlated** with religion orthodox

Bioplots: Length

Short arrows	don’t contribute much to either PC
Long arrows (pointing along an axis origin)	contribute a lot to one PC, but not the other
Long diagonal arrows	contribute a lot to both PCs

Bioplots: Length

Short arrows	don't contribute much to either PC
Long arrows (pointing along an axis origin)	contribute a lot to one PC, but not the other
Long diagonal arrows	contribute a lot to both PCs

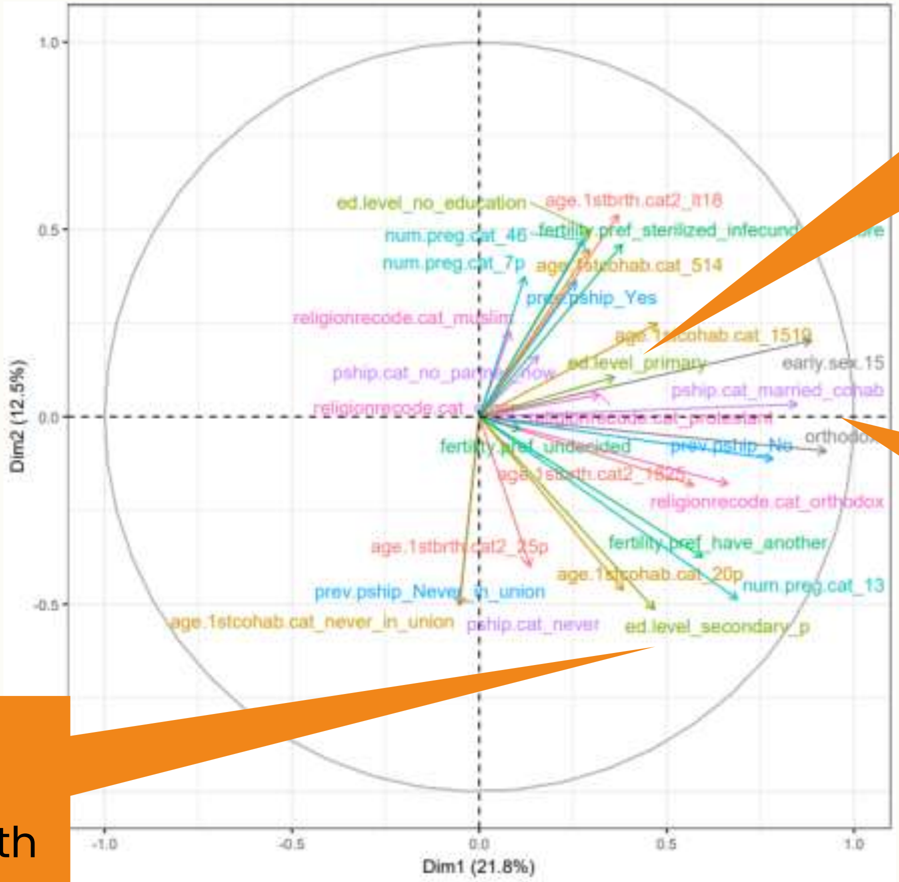
Is one PC more helpful in explaining variation than another?

Bioplots: Length

Short arrows	don’t contribute much to either PC
Long arrows (pointing along an axis origin)	contribute a lot to one PC, but not the other
Long diagonal arrows	contribute a lot to both PCs

The length of a PC does not imply this variable is more predictive of an outcome. We are **NOT** considering outcomes at all in PCA.

Example



Education level primary short and therefore not very informative to either PC1 or PC2

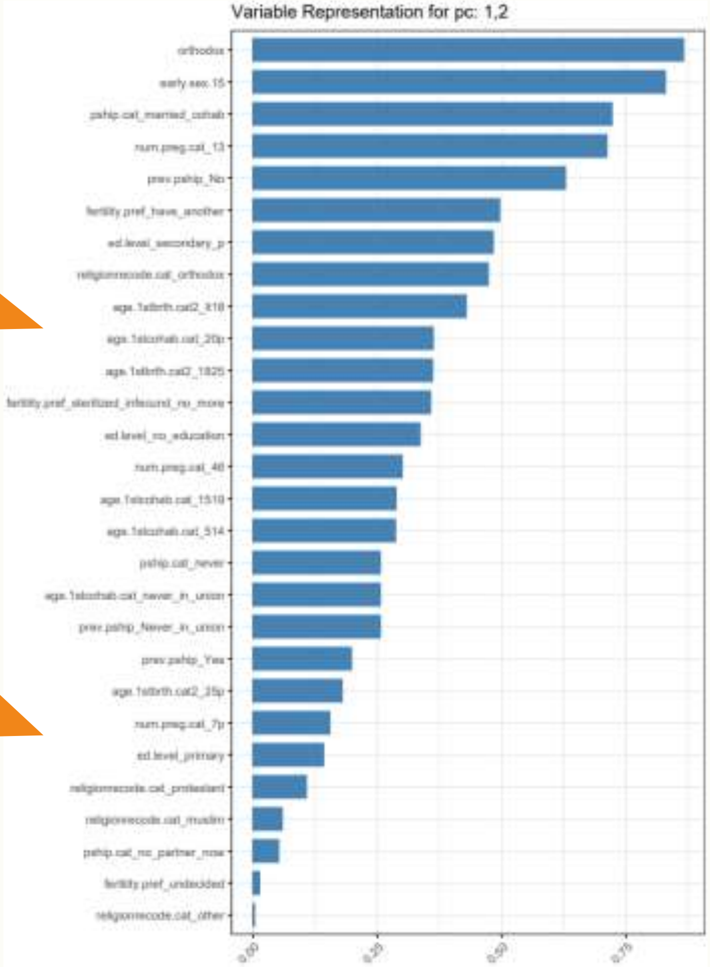
Partnership.married_cohab loads highly in one PC1 but not the other

Secondary and higher education loads high for both PCs

quality of representation barplot:

show how much a variable contributes to two PCs, as measured by cos2

Example



The larger the bars are, the closer they are to the circumference of the circle in the biplot

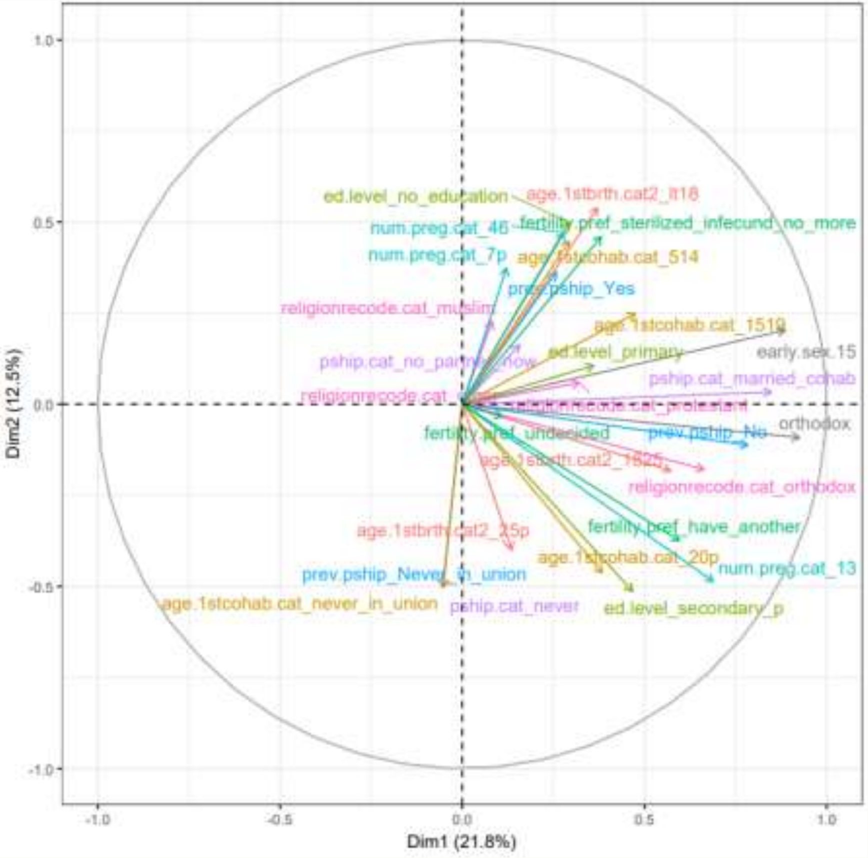
A high cos2 indicates the variable is well represented in the PCs (i.e., high loadings/long diagonal lines)

A low cos2 indicates that the variable is not well represented by the PCs (i.e., low loadings/short lines)

Categorical variables with multiple response categories

Interpret all arrows together:

- Do all the arrows for that variable point in the same direction?
- Do all the levels of the variables have the same length?
- Do all levels point in the same direction as other variables?

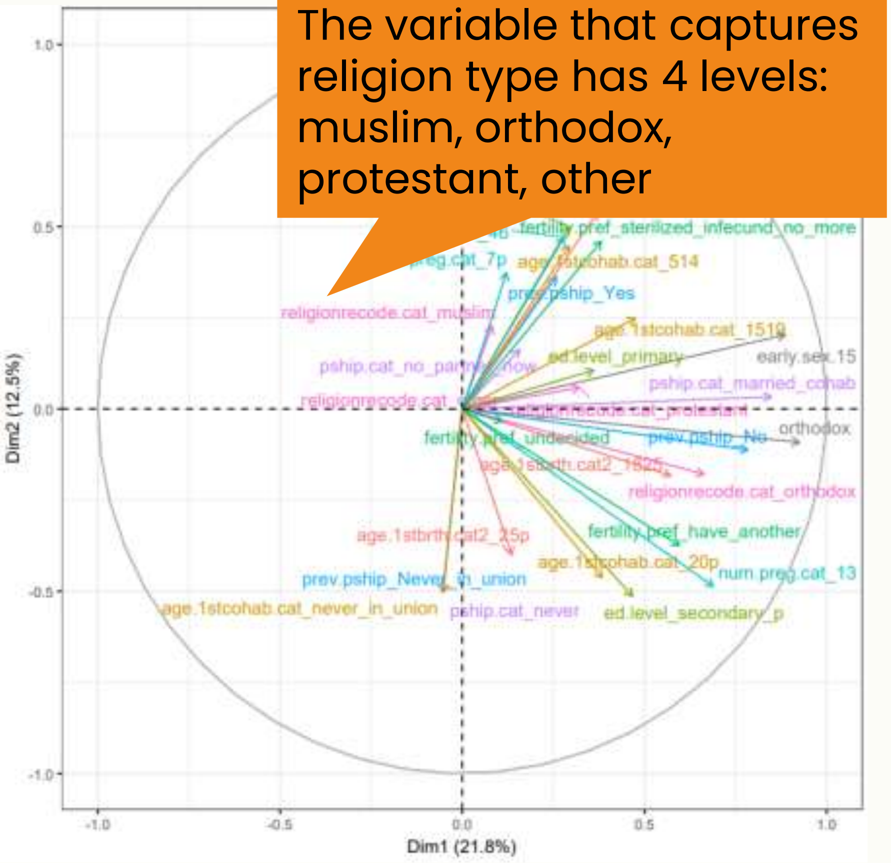


Categorical variables with multiple response categories

Interpret all arrows together:

- Do all the arrows for that variable point in the same direction?
- Do all the levels of the variables have the same length?
- Do all levels point in the same direction as other variables?

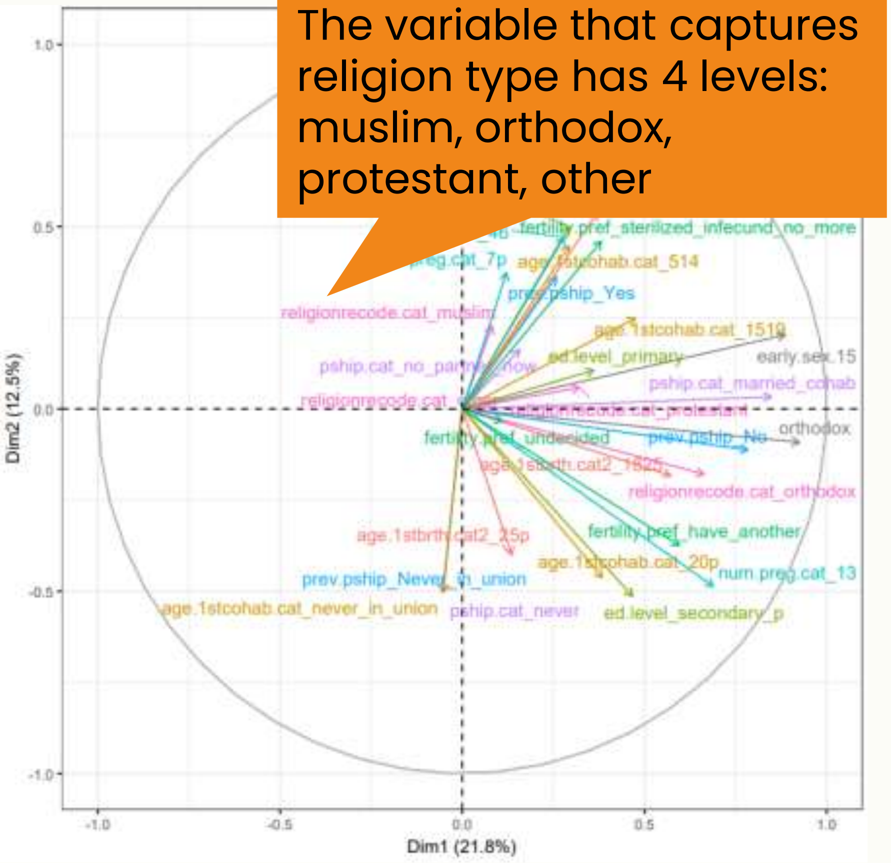
The variable that captures religion type has 4 levels: muslim, orthodox, protestant, other



Categorical variables with multiple response categories

Interpret all arrows together:

- Do all the arrows for that variable point in the same direction?
- Do all the levels of the variables have the same length?
- Do all levels point in the same direction as other variables?



The variable that captures religion type has 4 levels: muslim, orthodox, protestant, other

If at least one level of the variable is orthogonal to other variables, that is evidence in favor of keeping the variable for its explanatory power.

06

Group Activity

discussion

Let's go through an example together as a group, looking at scree plots, variable factor plots, and biplots!

homework

For your assigned domain, use the PCA output to decide which variables to keep.

https://uwashington.qualtrics.com/jfe/form/SV_79afsxNQg7A4ecm

Session survey