Steps in Data Management

(to be considered for EACH variable)

Primary variables

1. Create numeric dummy codes for variables containing *string data*
2. Code out missing data
3. Code in valid data

Creating NEW variables

1. Reverse or recode data (so it is potentially more logical)
2. Collapse categories within a variable (groups)
3. Make a new categorical variable by aggregating across multiple variables
4. Make a new quantitative variable by summing across multiple variables

Subset of observations

You have selected the variables (columns). Do you want to subset the observations (rows)?

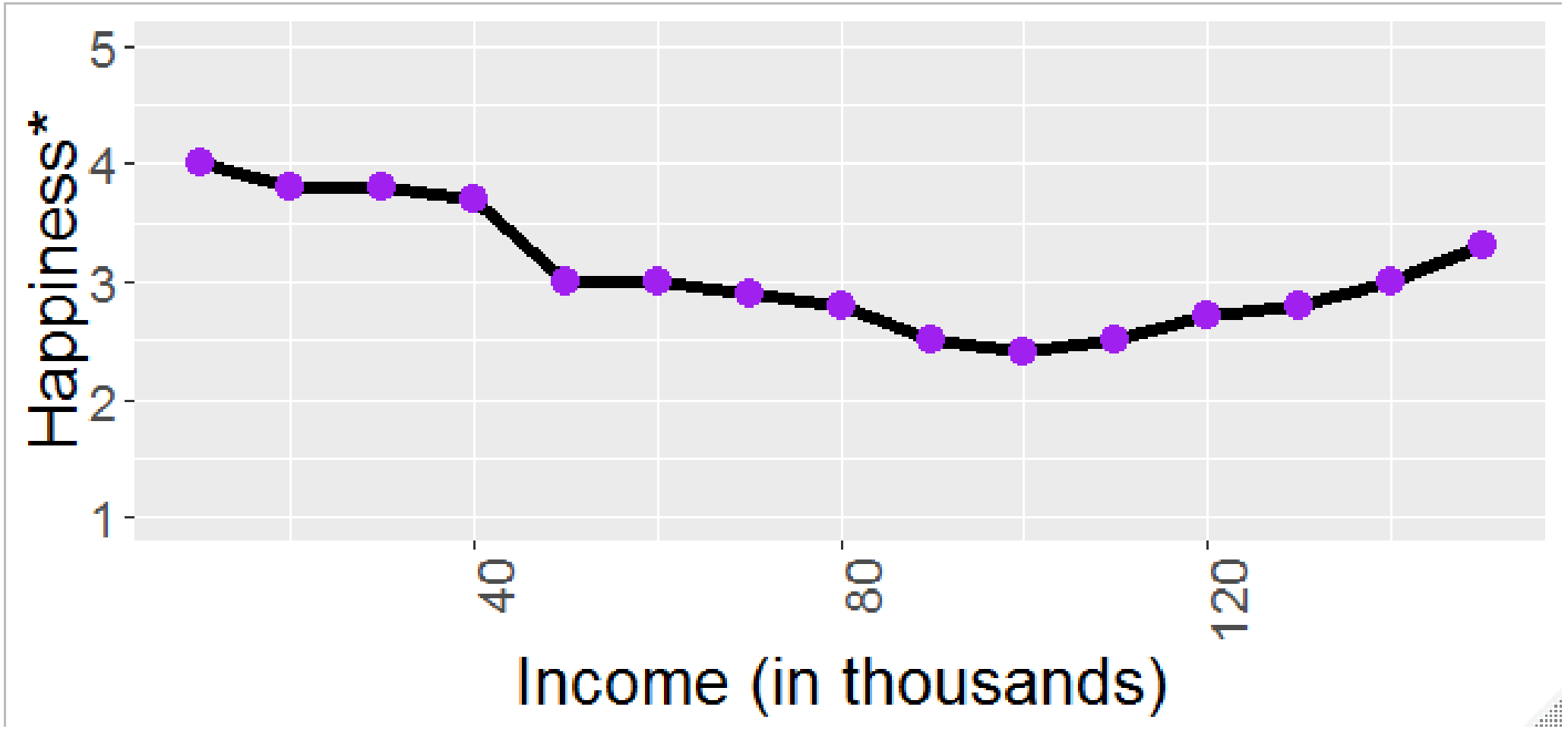
Sample Small Codebook

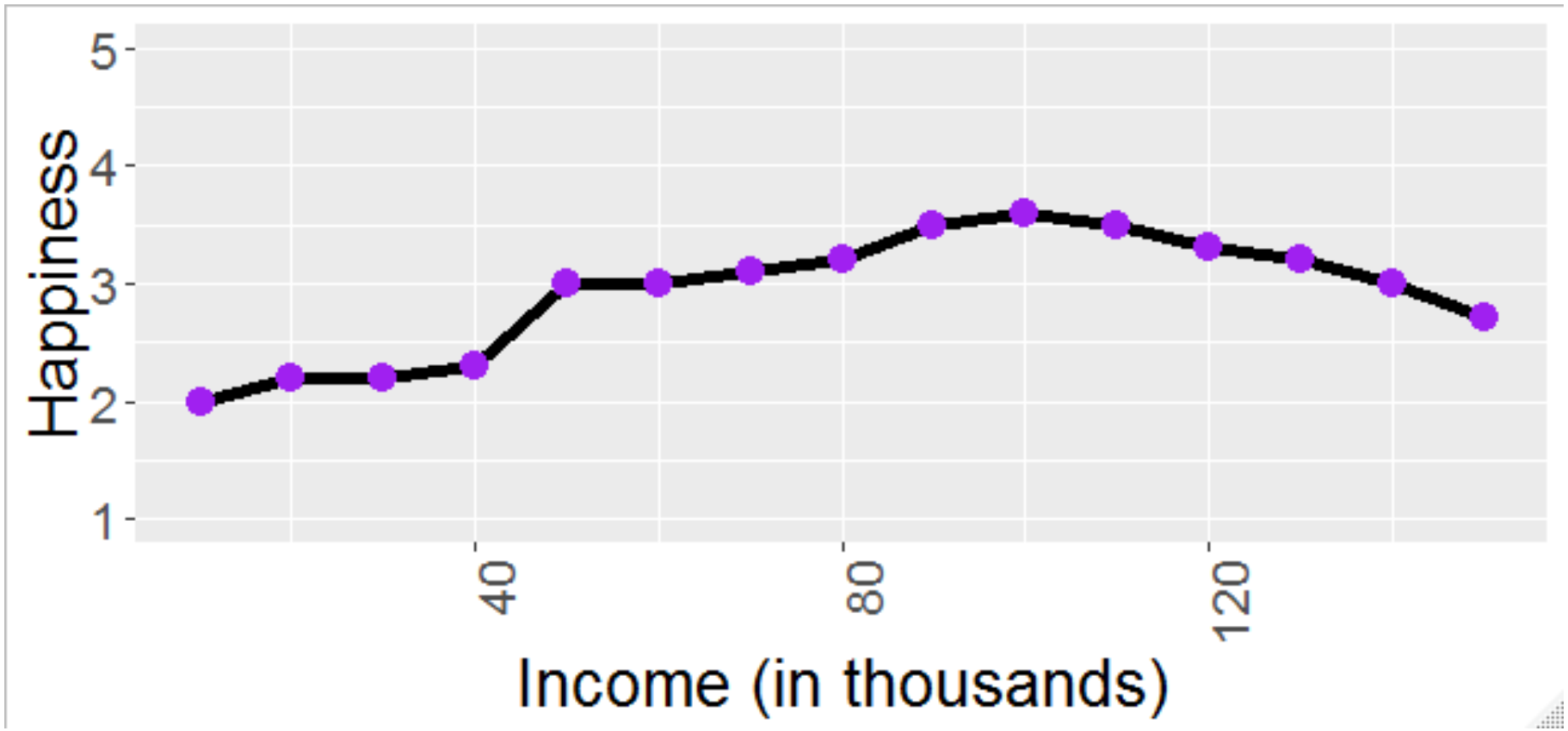
PatientID	Unique identifier for each patient
Sex	Male (M) and Female (F)
Weight	Numeric value in pounds, 999 denotes a refusal to answer
Height	Numeric value in inches, 999 denotes a resfusal to answer
Smoking	1(Yes), 2(No)
CollegeEducation	Any college education? 1(Yes), 2(No)
Years	Number of years of college education, skipped if 2 for CollegeEducation, 999 denotes refusal to answer
Happiness	1(Very Happy) 2 (Generally Happy) 3(So-so) 4(Generally Unhappy) 5(Very Unhappy) 998 (not asked), 999 (refused to answer)

Patient ID	Sex	Weight	Height	Smoking	CollegeEducation	Years	Happiness
A001	M	199	69	1	1	4	2
A002	M	210	70	1	1	4	2
A003	M	150	67	2	1	6	1
A004	F	145	64	2	1	8	3
A005	F	170	68	999	2		2
A006	M	120	59	1	2		4
…	…	…	…	…	…	…	
A967	F	999	999	2	1	3	5

Data Management Steps

- Sex:
  - Construct new variable, “SexID”:
    - 1 – if participant is Female
    - 0 – if participant is Male
- Weight
  - Code out missing data
  - That is, have 999 be read as missing
- Height
  - Code out missing data
  - That is, have 999 be read as missing
- Smoking:
  - Code out missing data
  - Have 999 be read as missing
- College Education
  - No data management necessary
- Years
  - Code in valid data
  - Have missing values be replaced with “0”.


Happiness*
5
4
3
2
1
40
80
120
Income (in thousands)


Happiness
5
4
3
2
1
40
80
120
Income (in thousands)

Sample Small Codebook 2

PatientID	Unique identifier for each patient
SocPhob	1 (Yes), 0(No)
Gad	1 (Yes), 0(No)
Panic	1 (Yes), 0(No)
Agora	1 (Yes), 0(No)
Ocd	1 (Yes), 0(No)

Patient ID	SocPhob	Gad	Panic	Agora	Ocd	
A001	1	0	1	0	0	
A002	1	1	0	0	1	
A003	0	0	0	0	0	
A004	0	0	0	0	0	
A005	0	0	0	0	0	
A006	0	0	0	0	0	
…	…	…	…	…	…	
A967	0	1	0	0	0	

Patient ID	SocPhob	Gad	Panic	Agora	Ocd	Anxiety
A001	1	0	1	0	0	1
A002	1	1	0	0	1	1
A003	0	0	0	0	0	0
A004	0	0	0	0	0	0
A005	0	0	0	0	0	0
A006	0	0	0	0	0	0
…	…	…	…	…	…	
A967	0	1	0	0	0	1

Patient ID	SocPhob	Gad	Panic	Agora	Ocd	AnxietyTotal
A001	1	0	1	0	0	2
A002	1	1	0	0	1	3
A003	0	0	0	0	0	0
A004	0	0	0	0	0	0
A005	0	0	0	0	0	0
A006	0	0	0	0	0	0
…	…	…	…	…	…	
A967	0	1	0	0	0	1

Sample Small Codebook 3

PatientID	Unique identifier for each patient
Highest Level of Education	7th grade, 8th grade, 9th grade, 10th grade, 11th grade, High School, Some College, Associates Degree, Bachelors Degree, MBA, MS, MD, PhD

Patient ID	Highest Level of Education	New Education Level
A001	8th grade	0
A002	MBA	2
A003	High School	1
A004	High School	1
A005	11th grade	0
A006	MS	2
…	…	…
A967	Bachelors	2

Original variable has 13 levels

We can collapse these categories into 3 groups, 0=less than HS, 1=High School, 2= Beyond HS

Example Code

1. Create numeric dummy codes for variables containing *string data*

   Make a new variable called ‘SexID’, where a 1 denotes a patient is female and a 0 denotes a patient is ‘Male’

   R: *data$SexID[data$Sex==“M”]<-0*
   data$SexID[data$Sex==“F”]<-1

   Stata: *gen SexID=.*
   replace SexID=1 if Sex==“F”
   replace SexID=0 if Sex==“M”

2. Code out missing data

For the variables Weight, Height, and Years, a 999 represents missing data, so we should code as missing.

R: *data$Weight[data$Weight==999]<-NA*
data$Height[data$Height==999]<-NA
data$Years[data$Years==999]<-NA

Stata: *replace Weight=. if Weight==999*
replace Height=. if Height==999
replace Years=. if Years==999

3. Code in valid data

Make Years=0 if missing

R: *data$Years[is.na(data$Years)]<-0*

Stata: *replace Years=0 if Years==.*