Introduction to Stata

Amy L. Johnson & Rebecca Gleit
Stanford University

Outline

Part 1: Data Organization

Part 2: Data Manipulation

Part 3: Self-Directed with Stata

Part 1: Data Organization

Data file: Friends.dta

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66
2	Math	Sophomore	Midwest,West	1	64
3	Sociology	Grad student	Northeast,Midwest,West	3	69
4	Sociology	Grad student	Northeast,Midwest,West	1	65
5	Sociology	Grad student	Northeast,West	4	65
6	Sociology	Grad student	Northeast,West	2	83
7		Co-term		0	77
8		Sophomore	South	1	88
9	Sociology of Education	Grad student	Northeast,West	1	63
10	Undeclared	Freshman	Midwest	.	38
11	Sociology!	Grad student	West	1	68
12	Sociology	Grad student	Midwest,West	1	70
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66
2	Math	Sophomore	Midwest,West	1	64
3	Sociology	Grad student	Northeast,Midwest,West	3	69
4	Sociology	Grad student	Northeast,Midwest,West	1	65
5	Sociology	Grad student	Northeast,West	4	65
6	Sociology	Grad student	Northeast,West	2	83
7		Co-term		0	77
8		Sophomore	South	1	88
9	Sociology of Education	Grad student	Northeast,West	1	63
10	Undeclared	Freshman	Midwest	.	38
11	Sociology!	Grad student	West	1	68
12	Sociology	Grad student	Midwest,West	1	70
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66

Each row is a different person

regions[1]	Northeast,West			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

regions[1]	Northeast,West

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

year_school[1]	3

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

regions[1]	Northeast,West

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

year_school[1]	3

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

height[1]	66

	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

String

regions[1]	Northeast,West			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

Numeric, with labels

year_school[1]	3			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

Numeric, without labels

height[1]	66			
	major	year_school	regions	siblings	height
1	Spanish	Junior	Northeast,West	1	66

Ways to Store Information

1. **String:**

Data are stored and appear as text

Ways to Store Information

1. **String:**

Data are stored and appear as text

2. Numeric

Ways to Store Information

1. **String:**

Data are stored and appear as text

2. Numeric

2A. **With labels:**

Data appear to be text, but are actually stored in the computer as numbers

Ways to Store Information

1. **String:**

Data are stored and appear as text

2. Numeric

2A. **With labels:**

Data appear to be text, but are actually stored in the computer as numbers

2B. **Without labels:**

Data are stored and appear as numbers

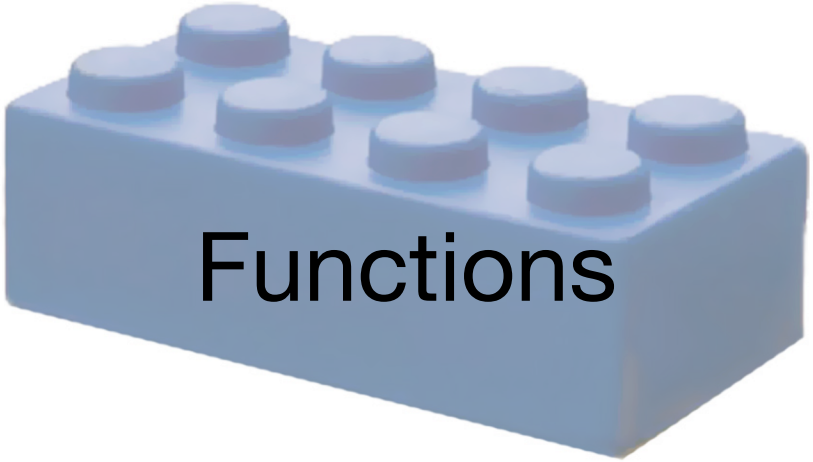
ACTIVITY: Determine the type of variable (string, numeric with labels, numeric without labels) for each variable.

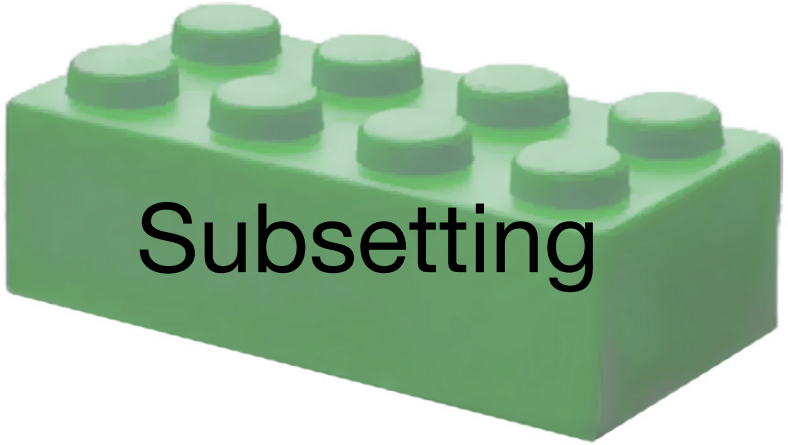
	major	year_school	regions	siblings	height	temp	F_C	cheese
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
7		Co-term		0	77	0	C	Gouda
8		Sophomore	South	1	88	.	.	
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya

Vars: 8 Order: Dataset Obs: 13

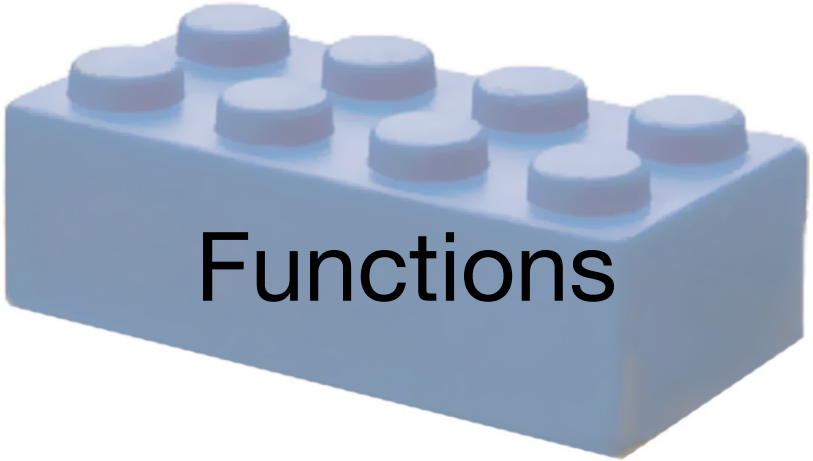
Part 2: Data Manipulation

The building blocks of data manipulation:

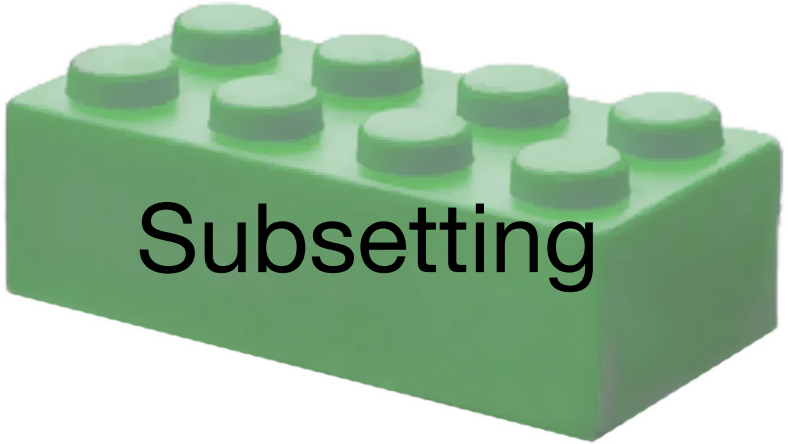




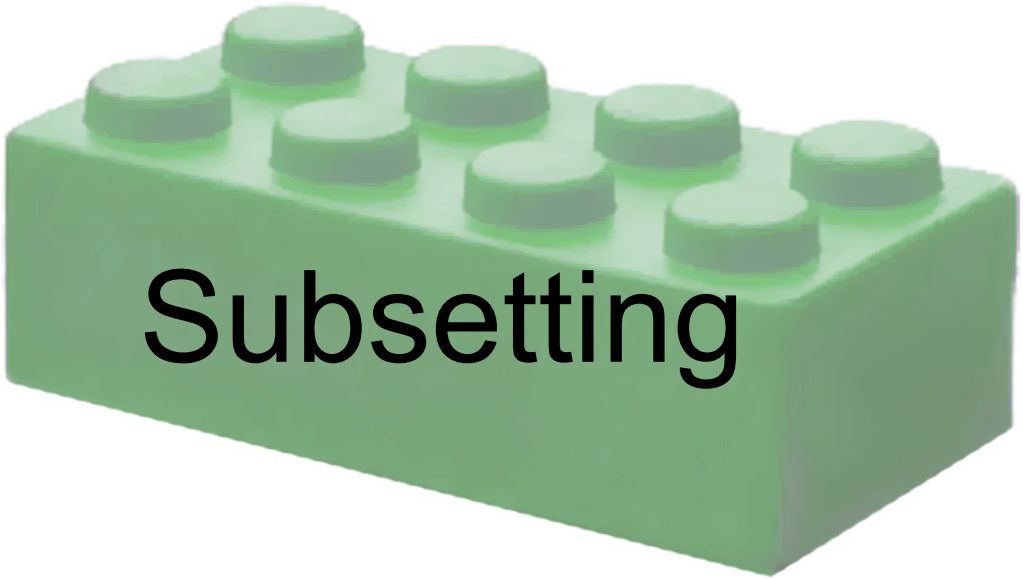
The building blocks of data manipulation:



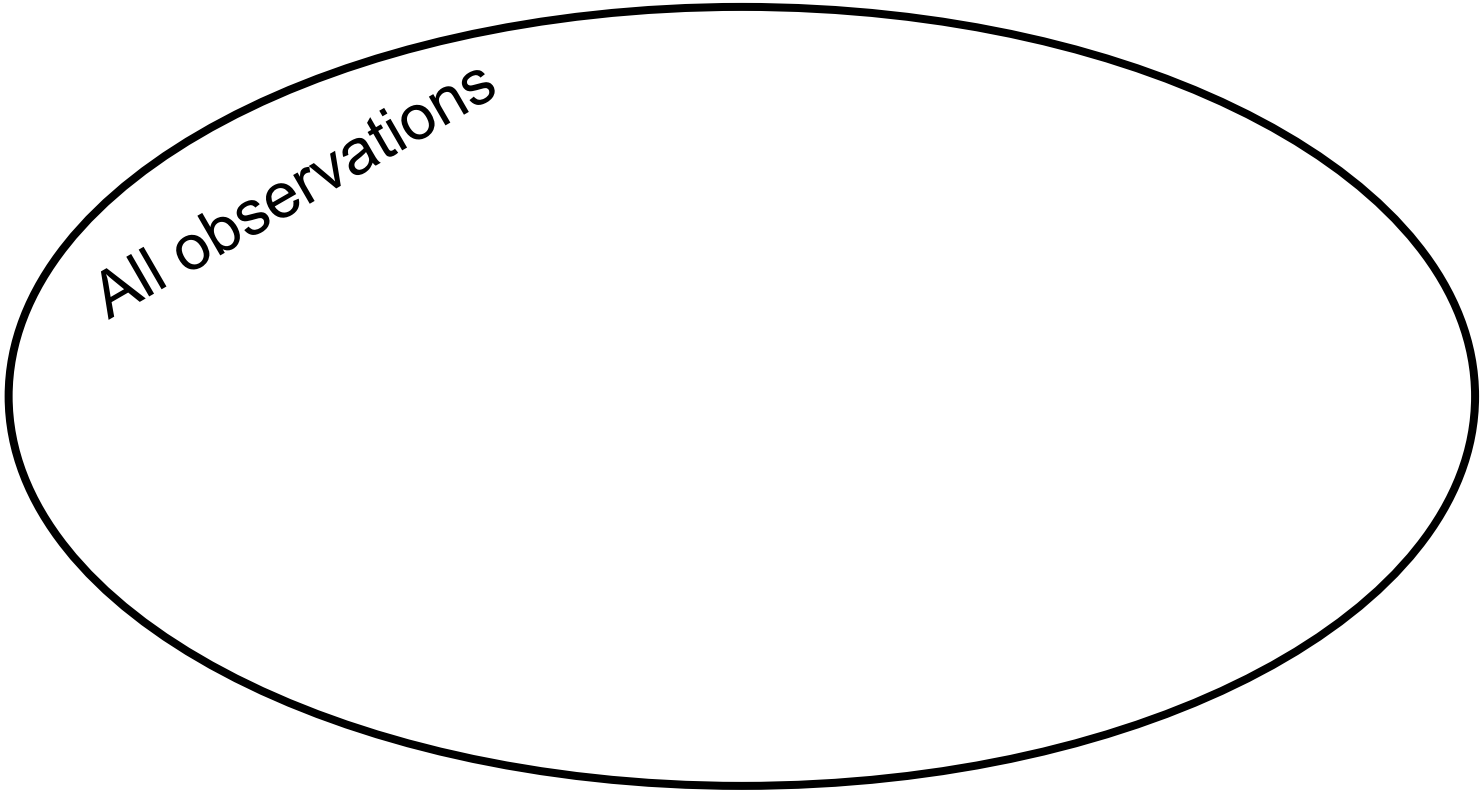
what we want to do

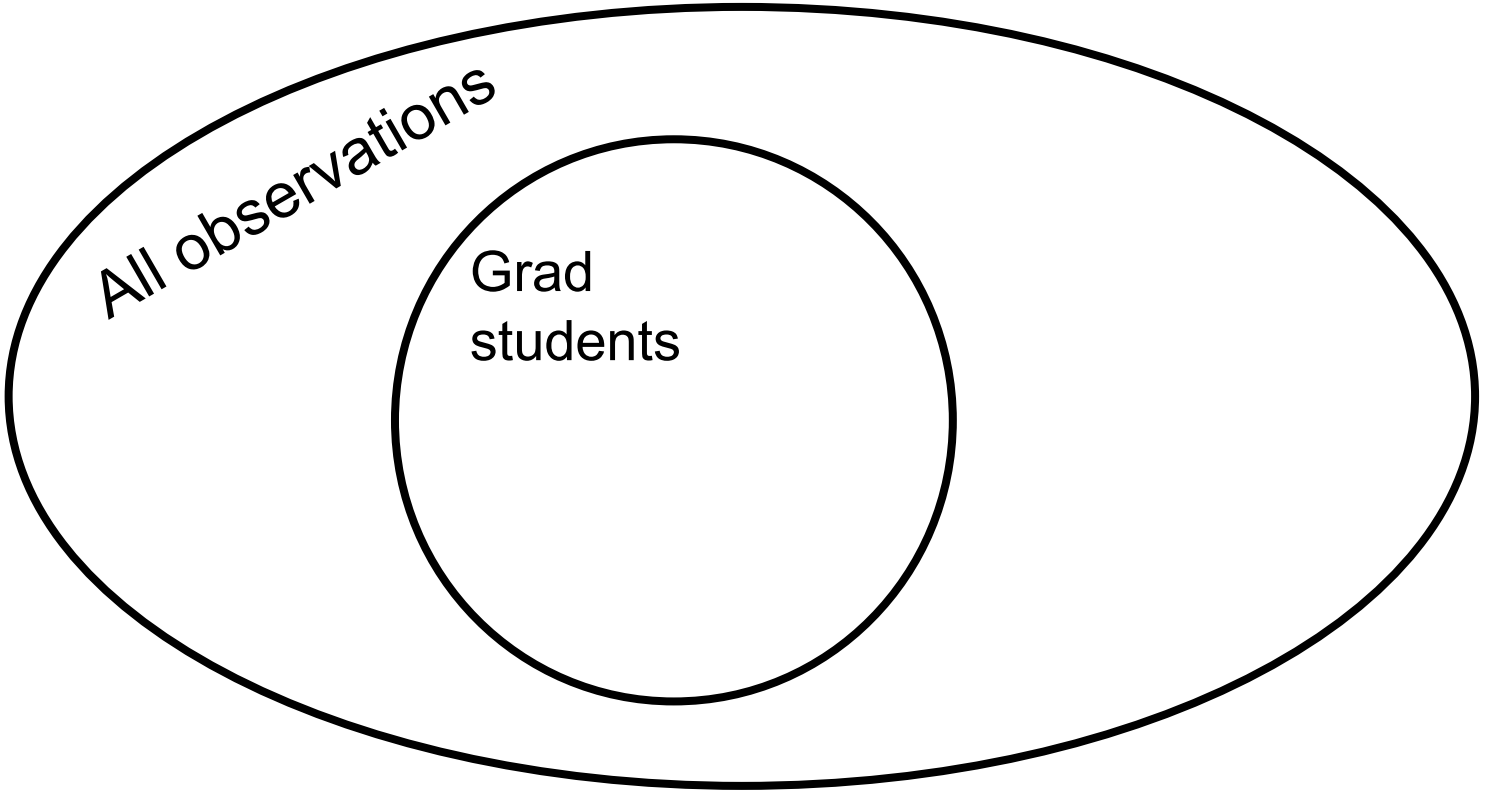


which observations we want to use the function on

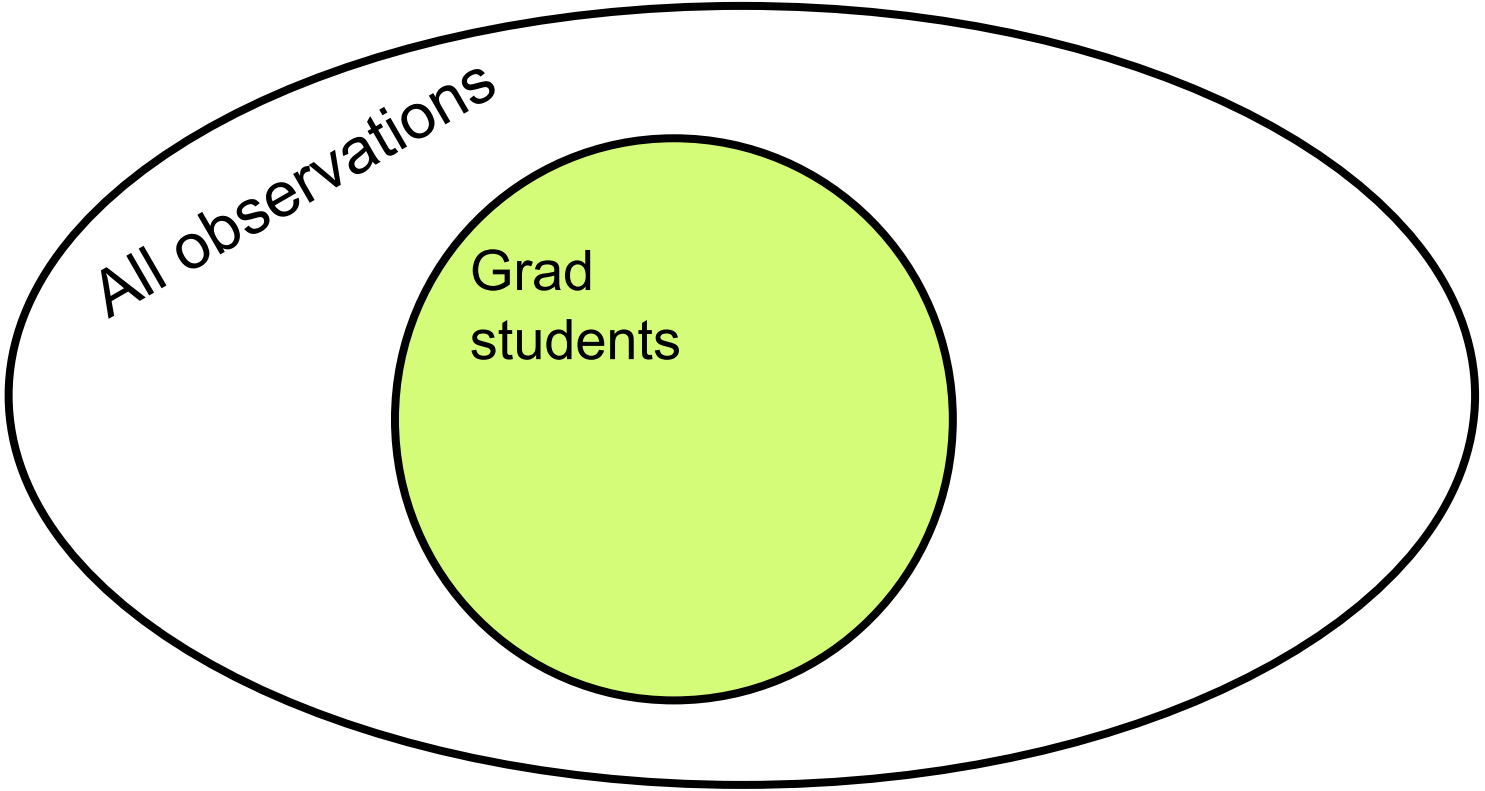

Subsetting

Using logical if-statements to subset

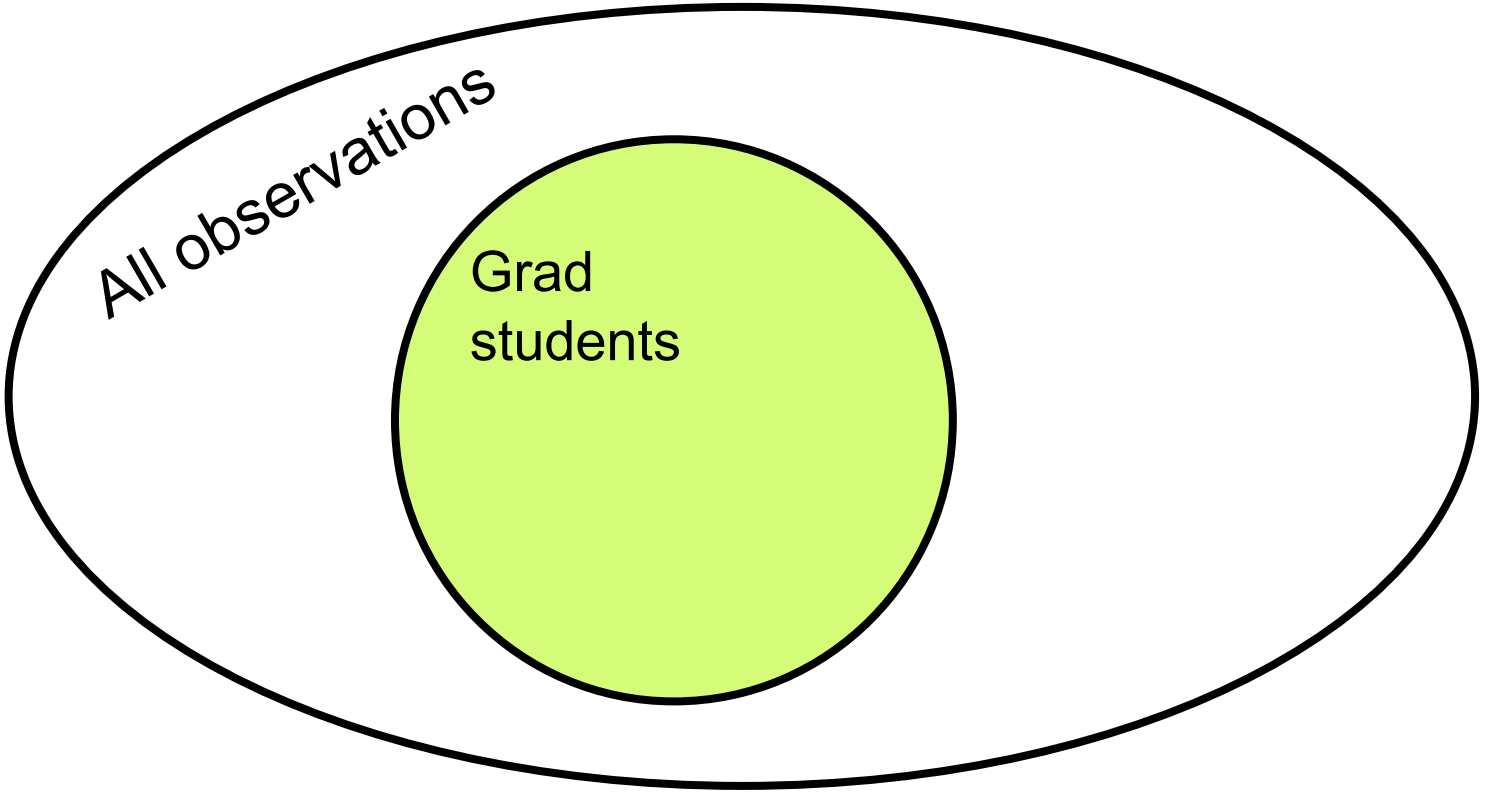



All observations
Grad
students

Look at people **if** they are a grad student

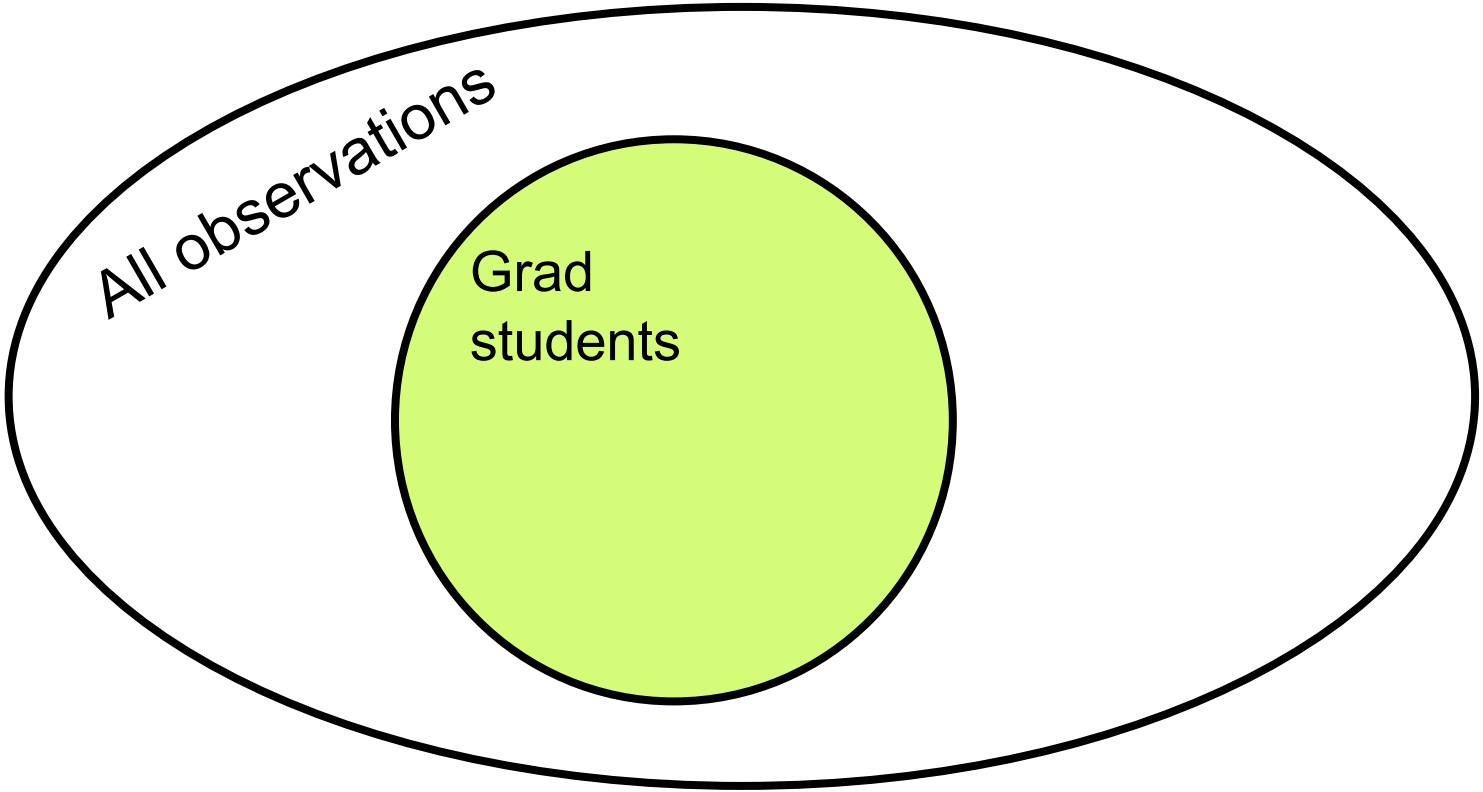


if year_school is equal to 6 ← (words)

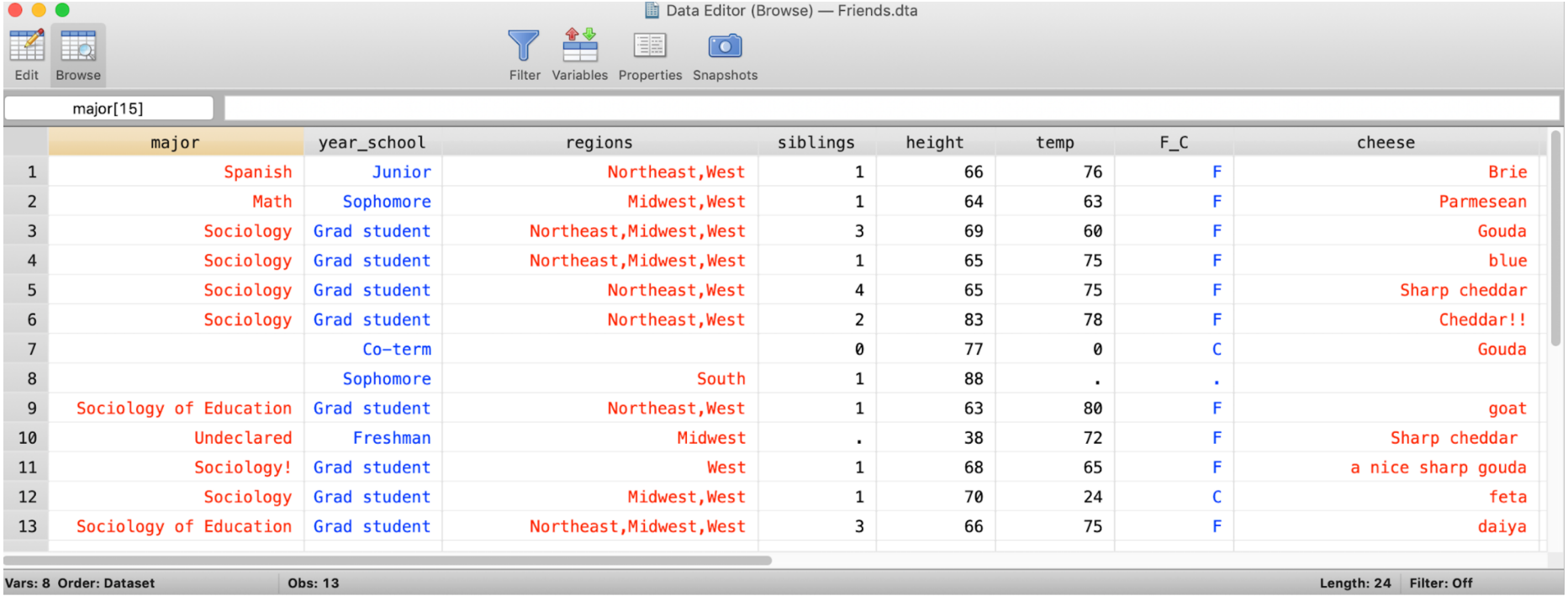


if year_school is equal to 6 ← (words)

```
if year_school==6
```
← (Stata syntax)

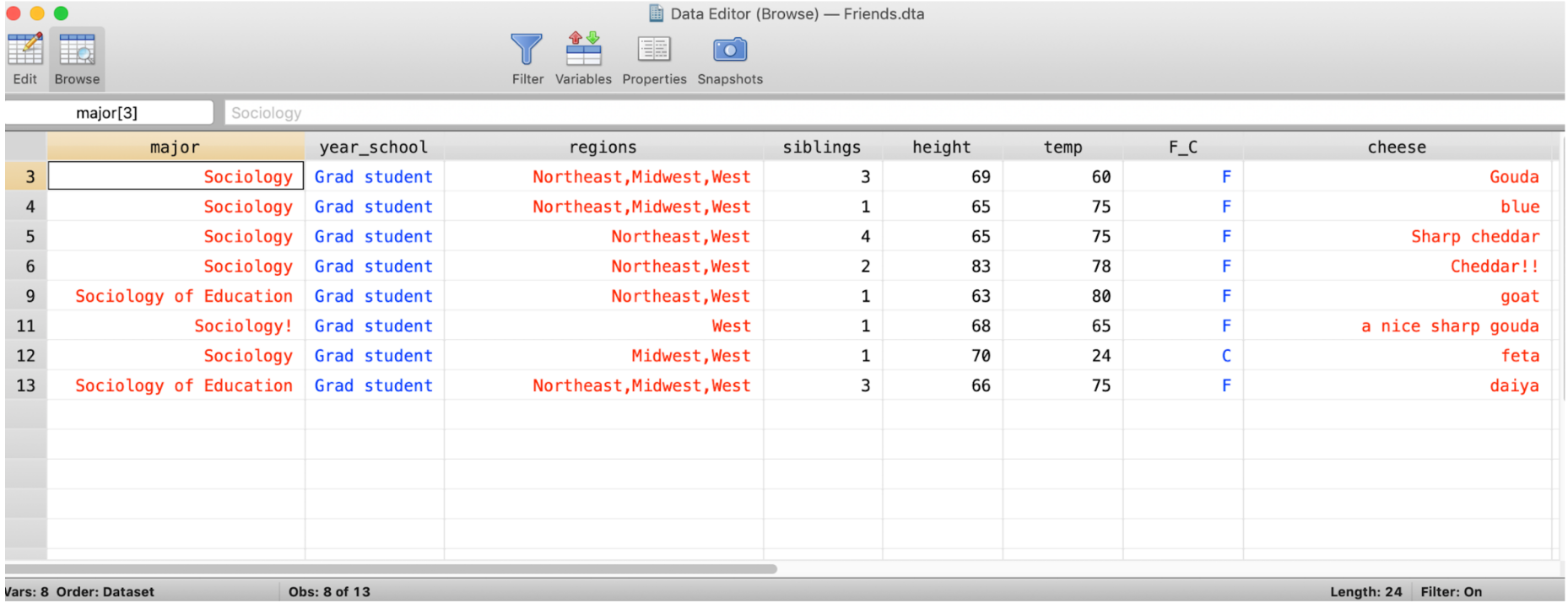


browse

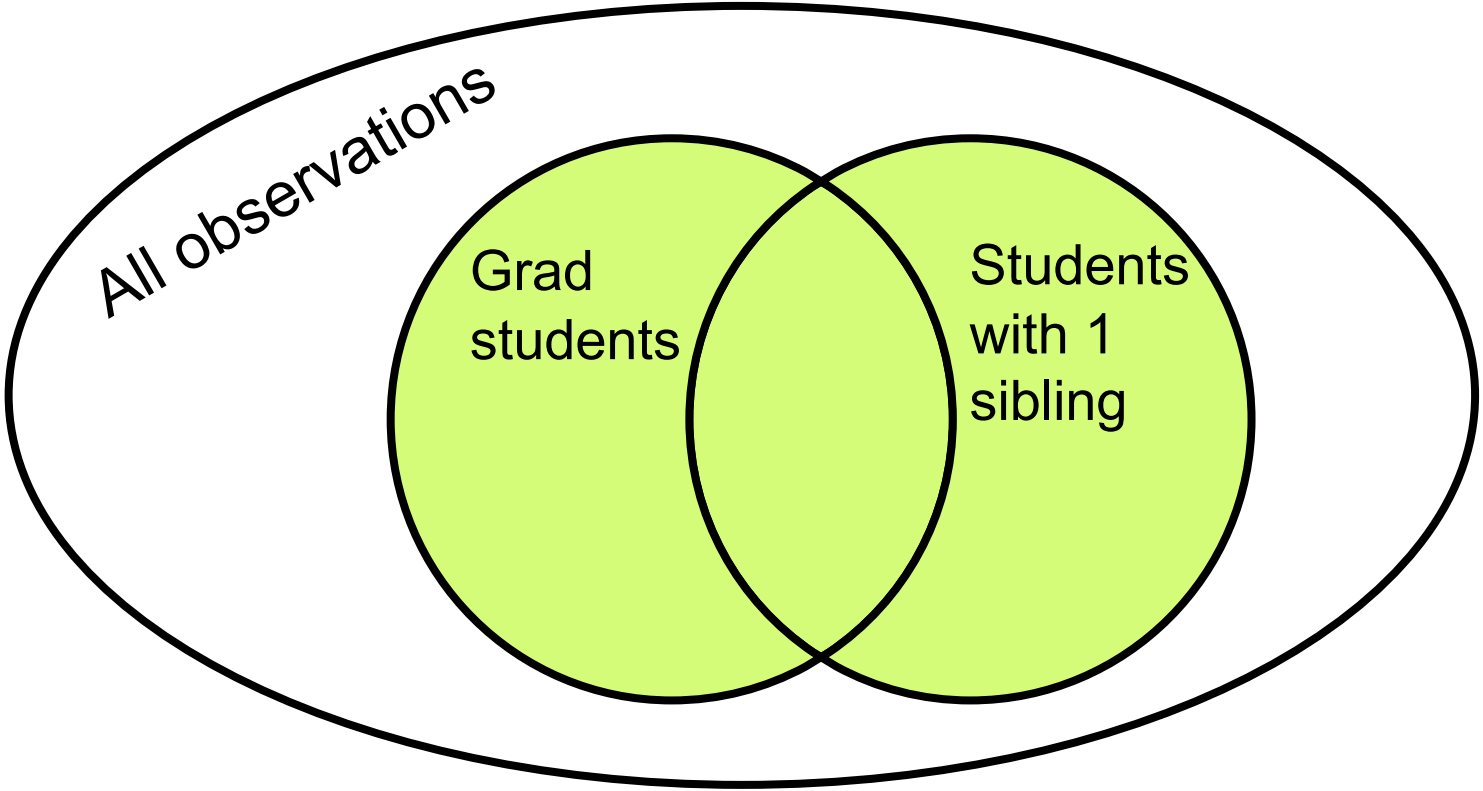


	major	year_school	regions	siblings	height	temp	F_C	cheese
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
7		Co-term		0	77	0	C	Gouda
8		Sophomore	South	1	88	.	.	
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya

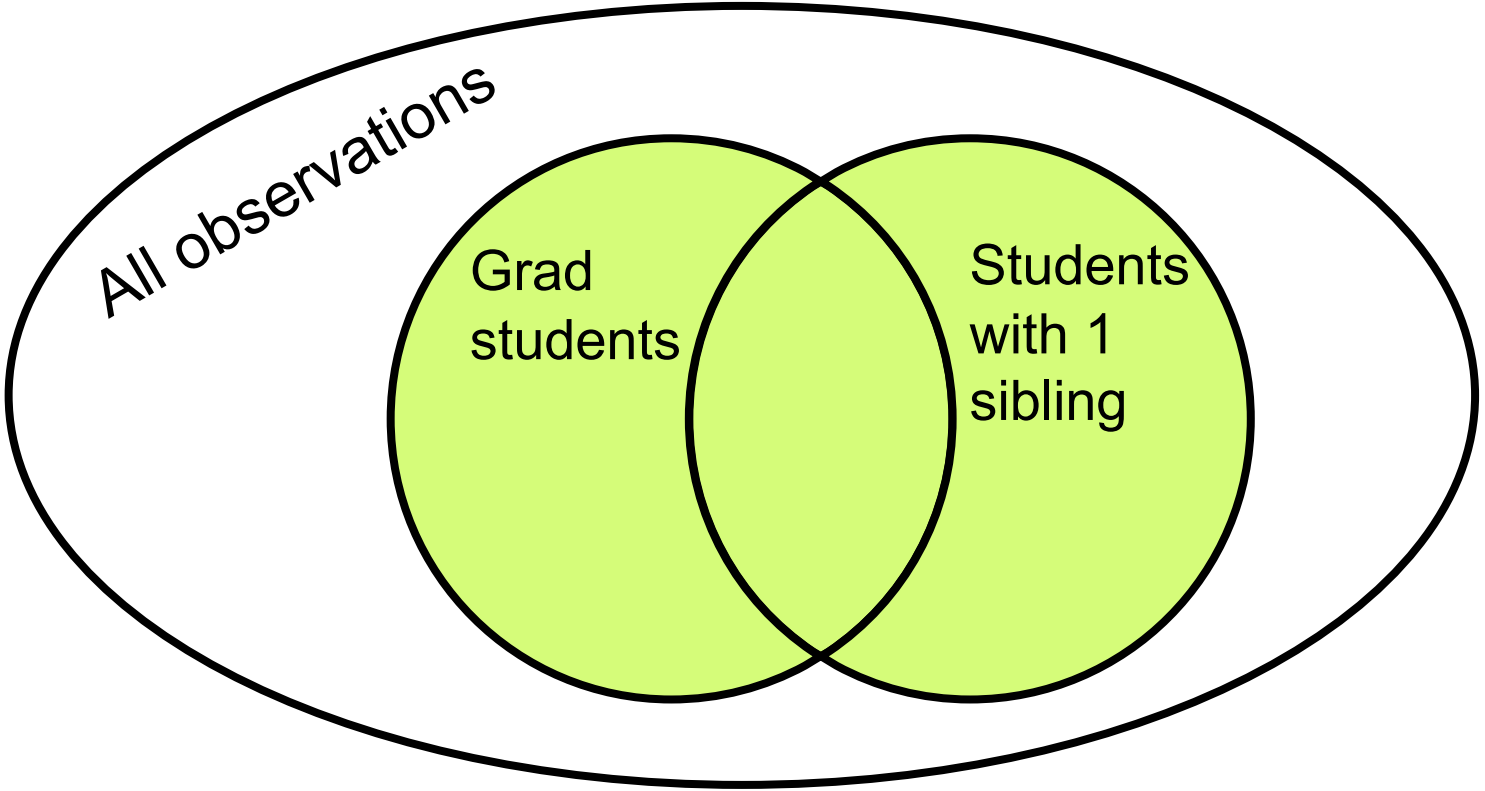
browse if year_school==6



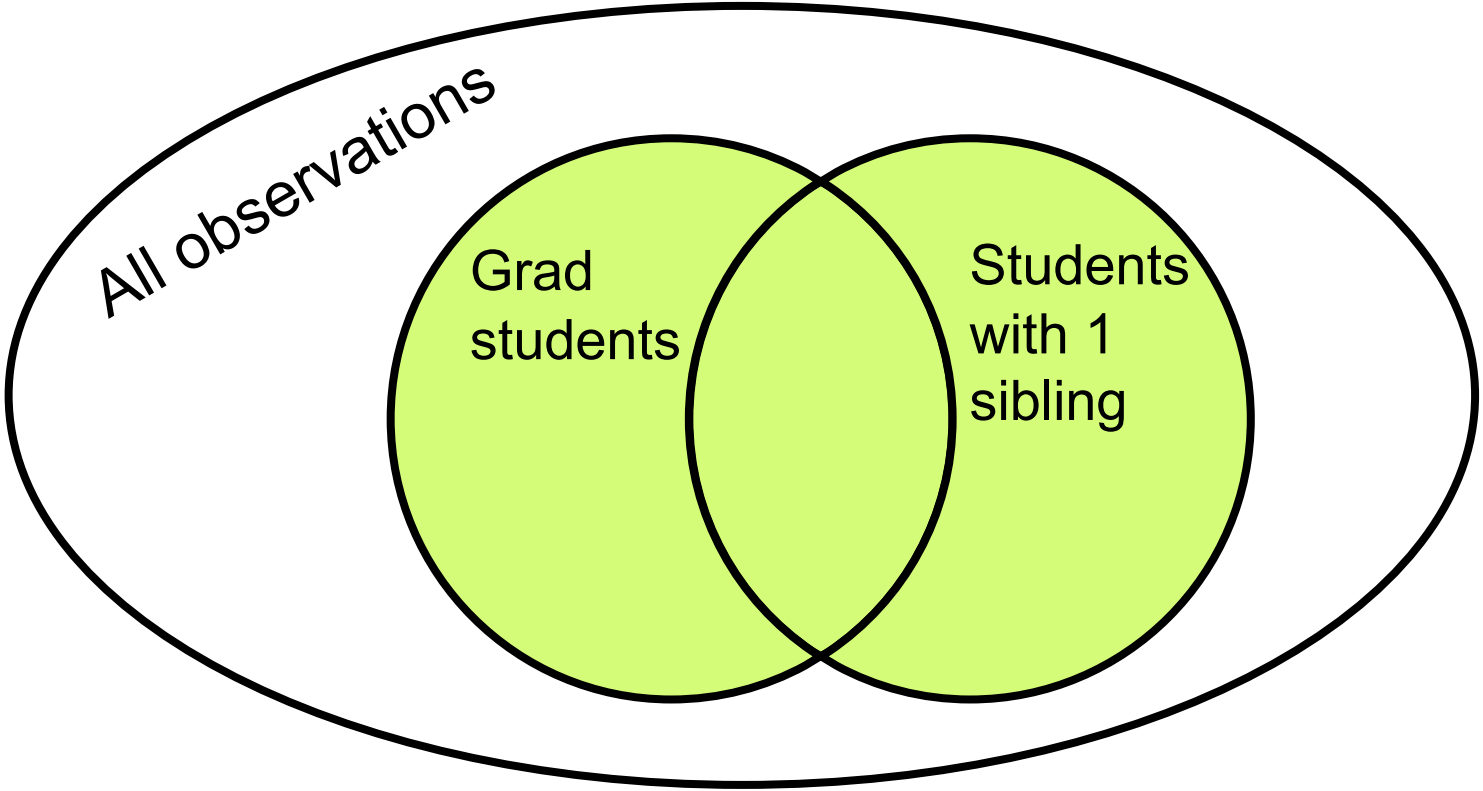
	major	year_school	regions	siblings	height	temp	F_C	cheese
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya


All observations
Grad students
Students with 1 sibling

Look at people **if** they are a grad student OR they have 1 sibling

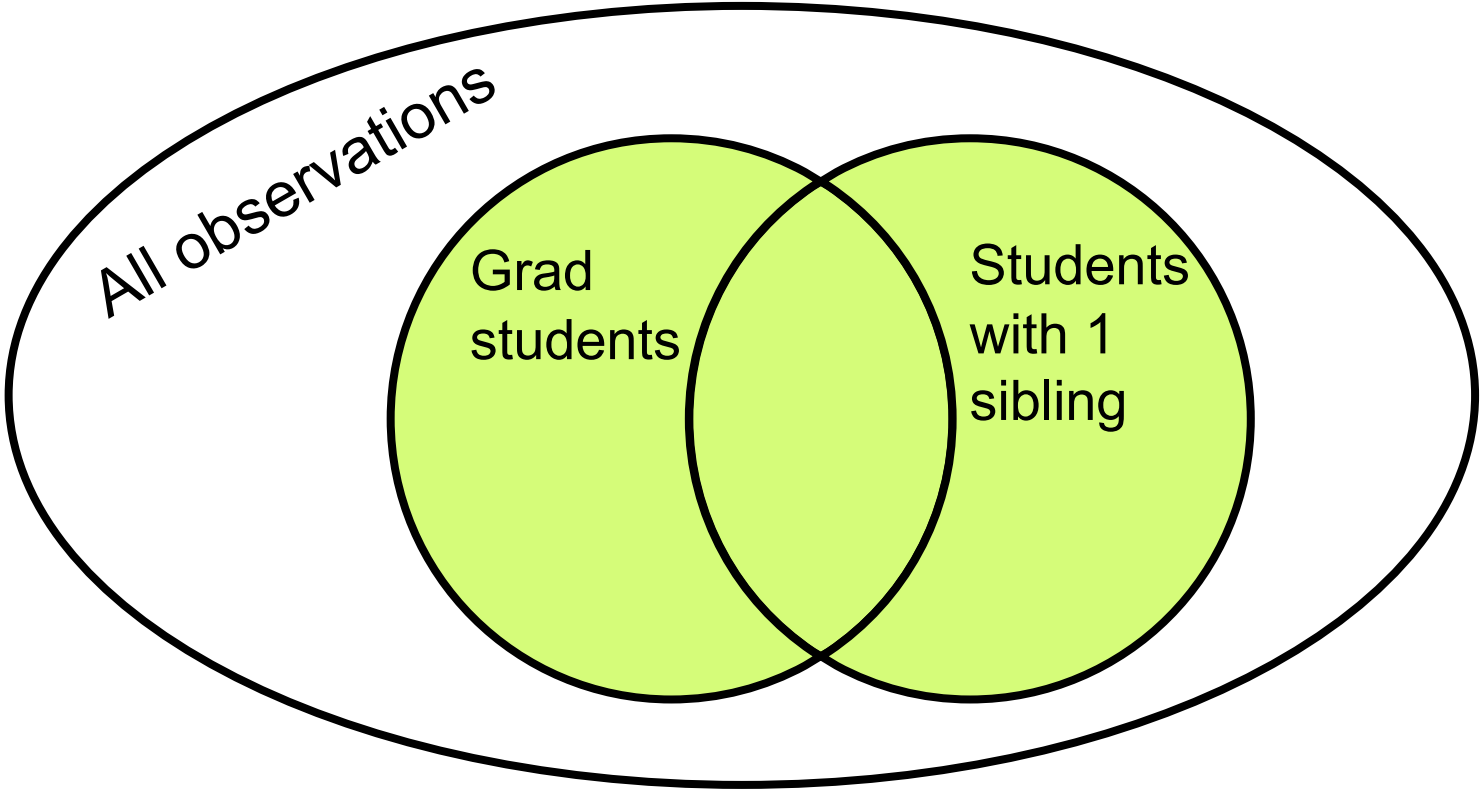


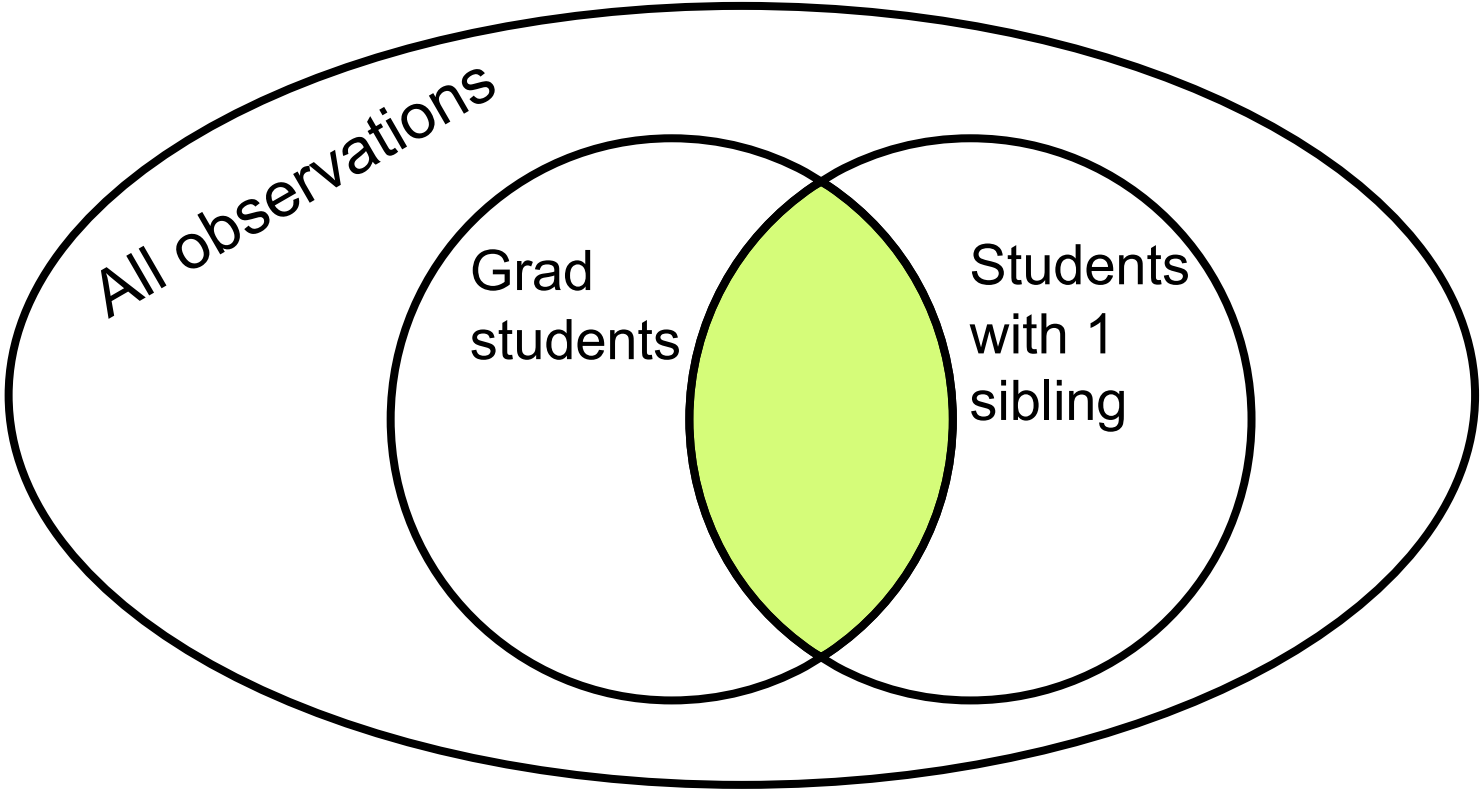
if year_school is equal to 6 OR siblings is equal to 1



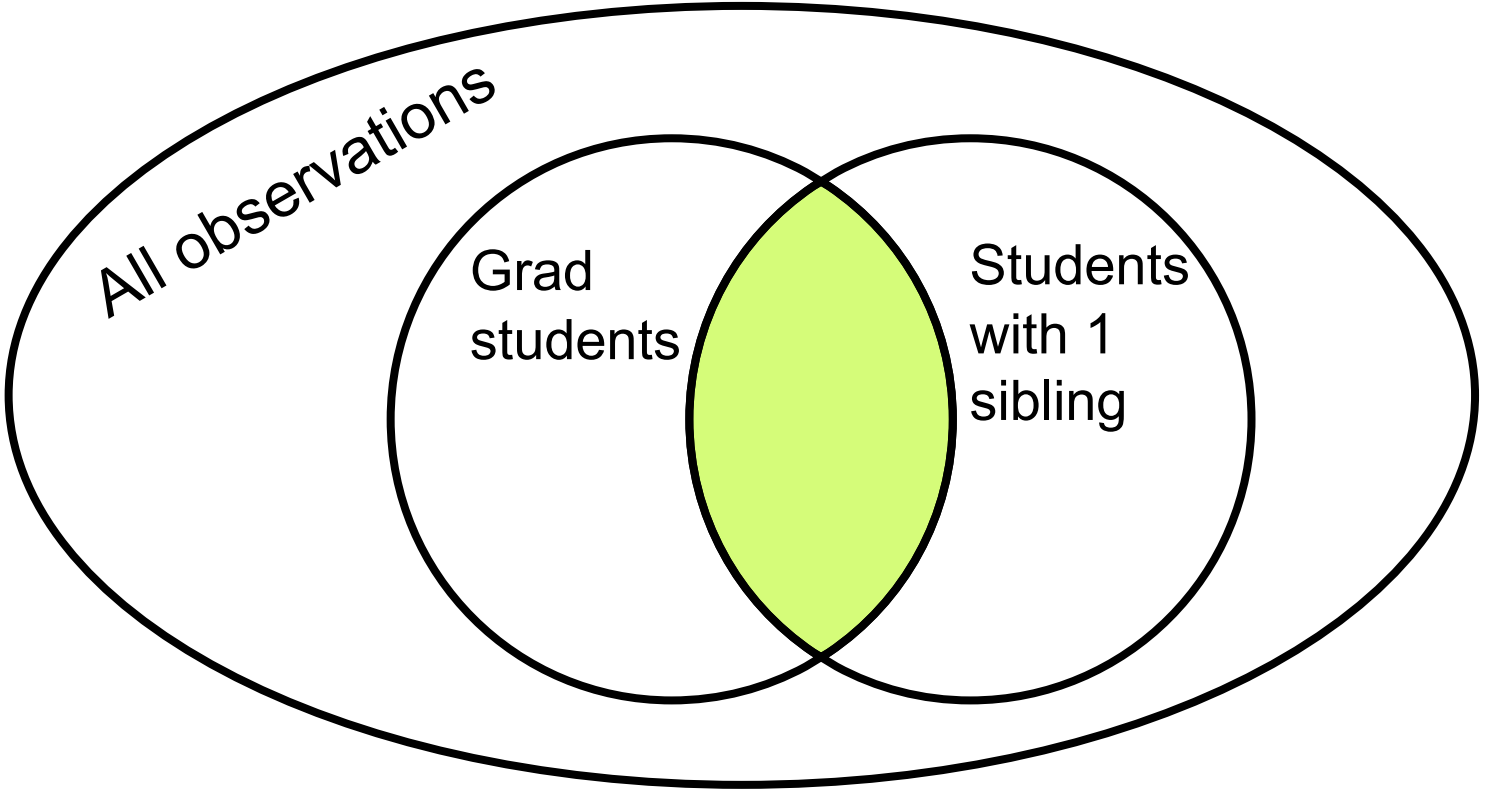
if year_school is equal to 6 OR siblings is equal to 1

```
if year_school==6 | siblings==1
```

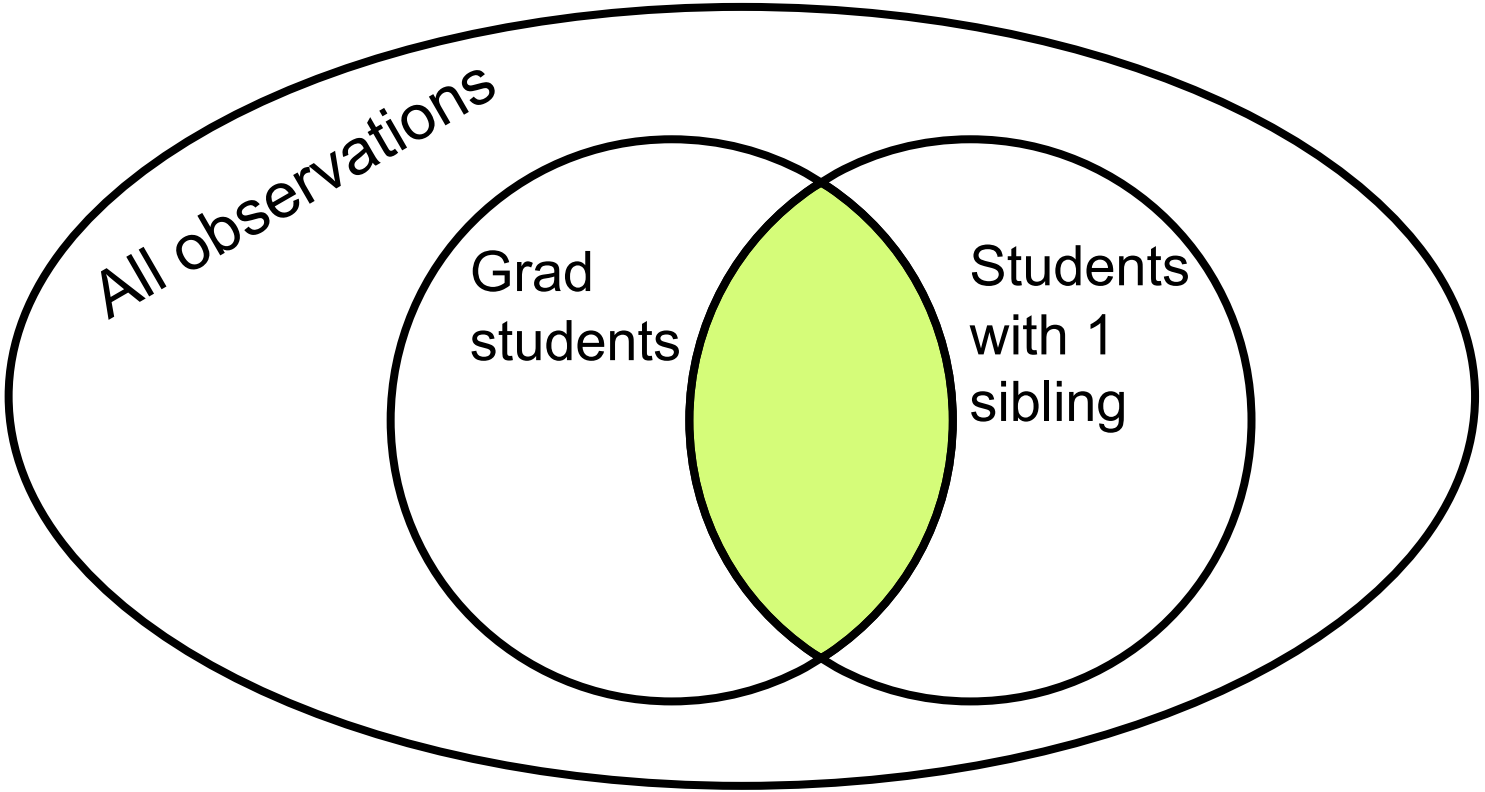



All observations
Grad students
Students with 1 sibling

Look at people **if** they are a grad student AND they have 1 sibling

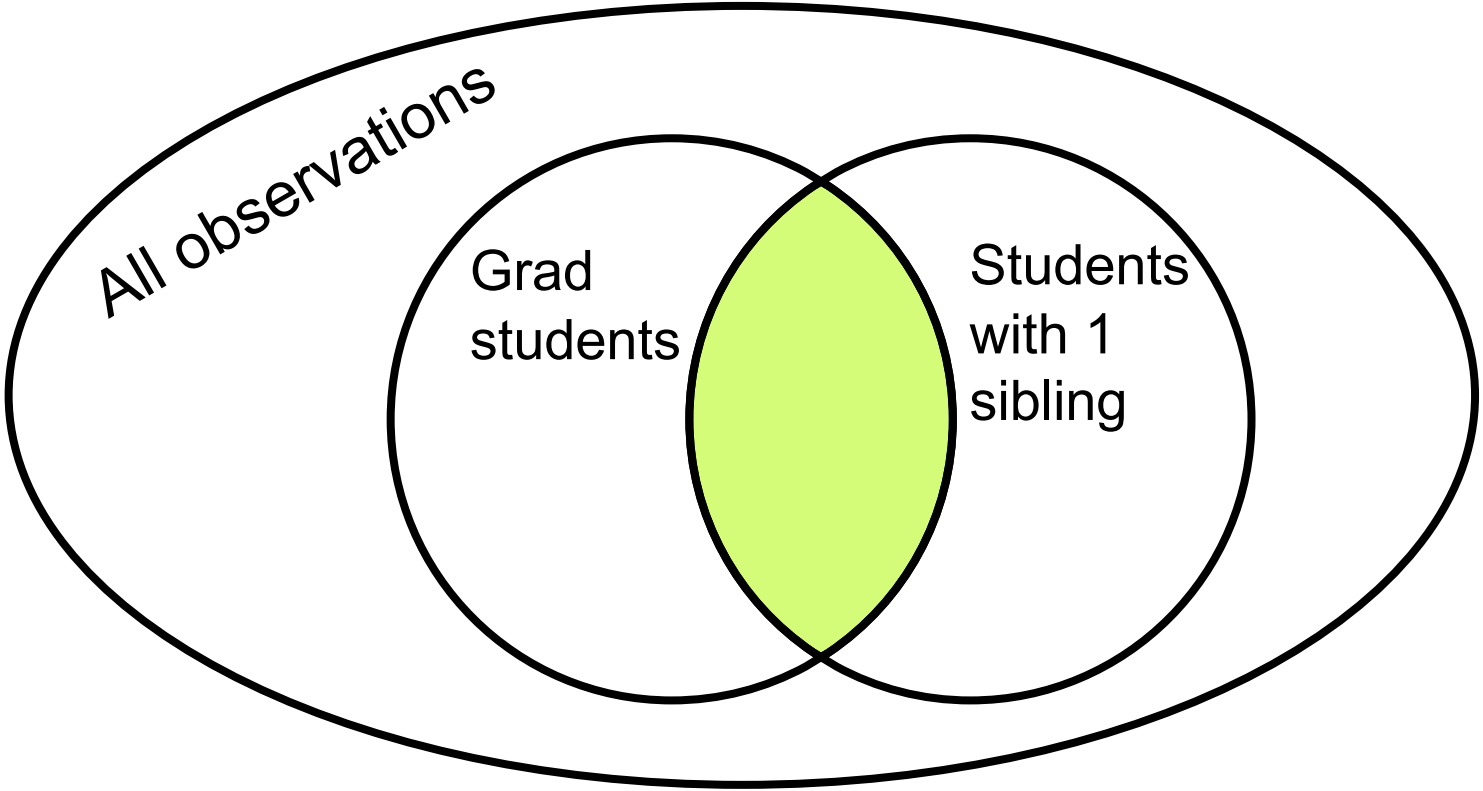


if year_school is equal to 6 AND siblings is equal to 1



if year_school is equal to 6 AND siblings is equal to 1

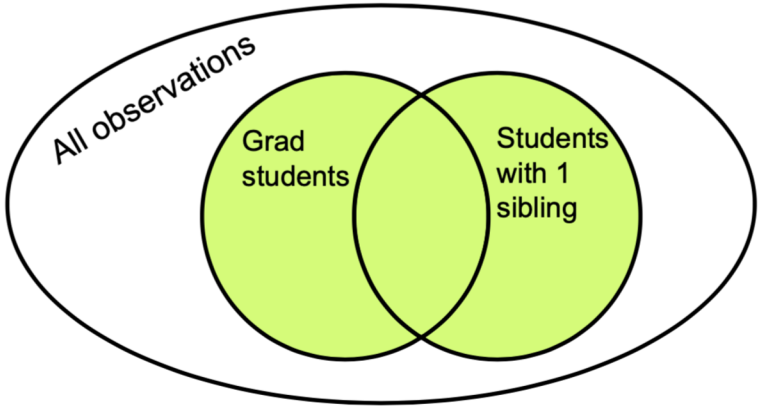
```
if year_school==6 & siblings==1
```



SUMMARY: Using logical statements to subset

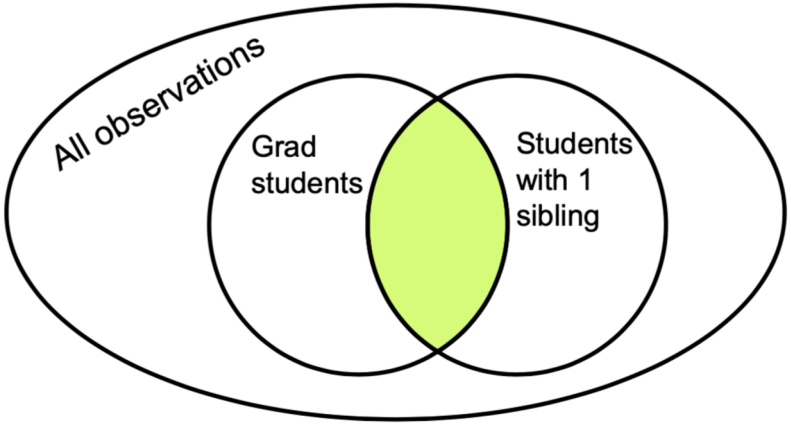
"OR": |

Must meet *at least 1* criteria



"AND": &

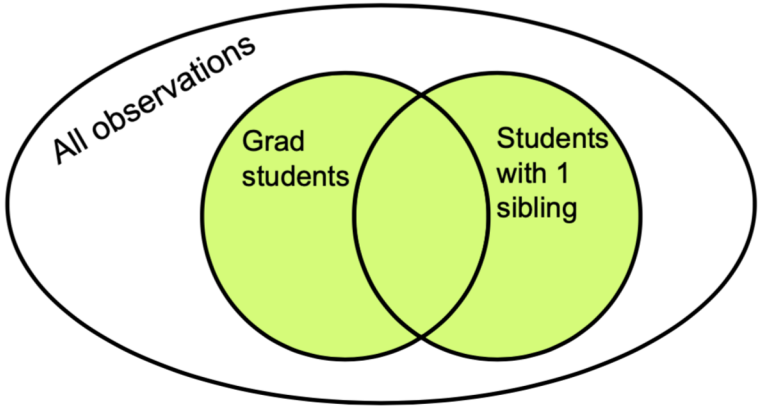
Must meet *all* criteria



SUMMARY: Using logical statements to subset

"OR": |

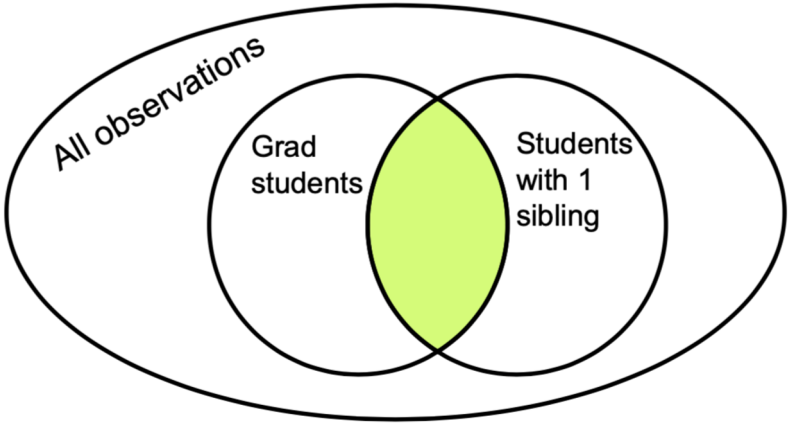
Must meet *at least 1* criteria



think: ***all*** *areas*

"AND": &

Must meet *all* criteria

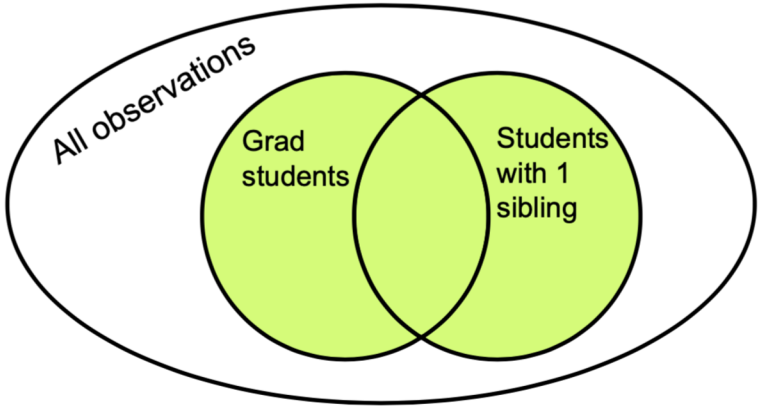


think: ***overlapping*** *areas*

ACTIVITY: “Practice subsetting observations” #1-13

“OR”: |

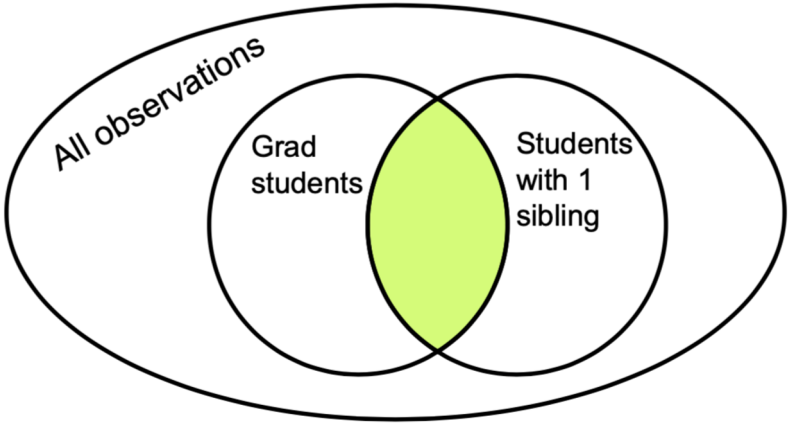
Must meet *at least 1* criteria



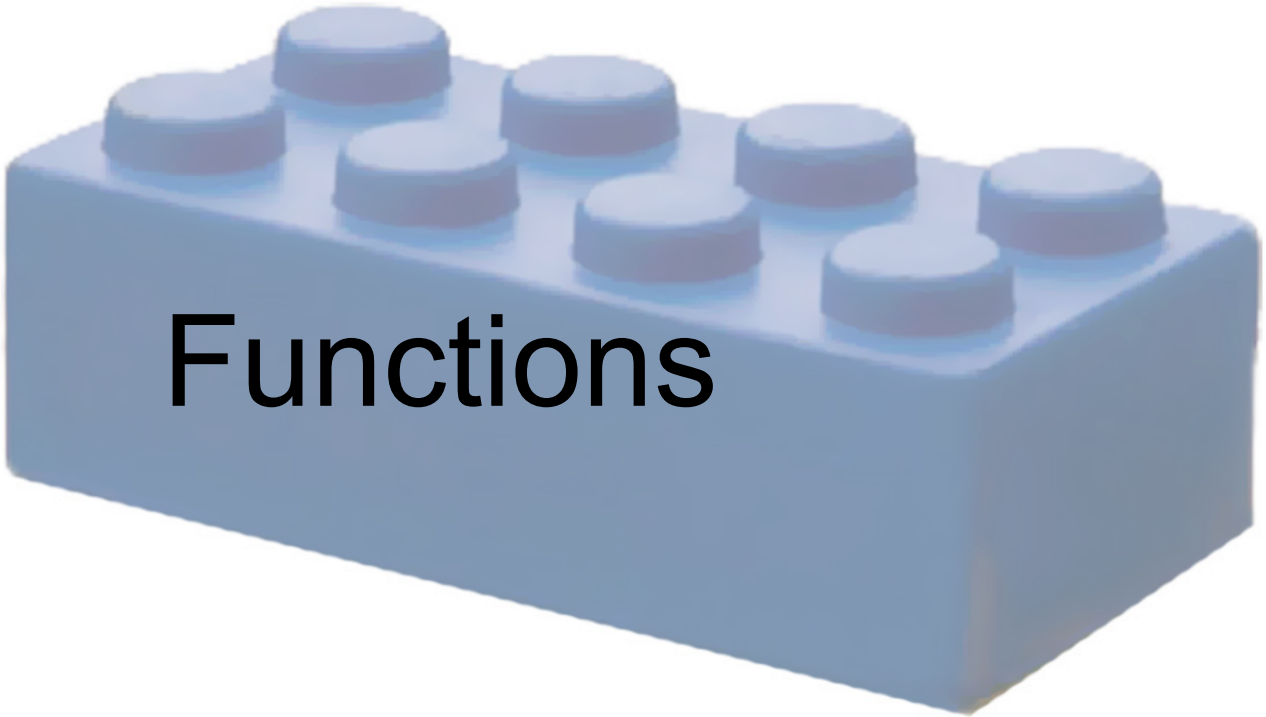
think: ***all*** *areas*

“AND”: &

Must meet *all* criteria



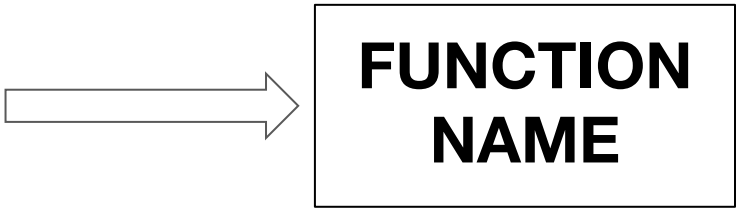
think: ***overlapping*** *areas*


Functions

Functions

Functions

Information included in input code



Functions


Information included in input code
FUNCTION NAME
Information included in the output

Example. We want to create a new variable that tells us whether someone is an only child. How would we do this?

Example. We want to create a new variable that tells us whether someone is an only child. How would we do this?

Answer: Use the variable *siblings* to create a different variable called *onlychild* where 1 = “is an only child” and 0 = “not an only child.”
(this is called a dummy variable)

To create new variables, we use two functions: **generate** and **replace**.

	major	year_school	regions	siblings	height	temp	F_C	cheese
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!
7		Co-term		0	77	0	C	Gouda
8		Sophomore	South	1	88	.	.	
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya

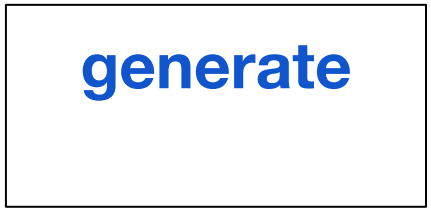
generate onlychild = 0

	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	0
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

replace onlychild = 1 if siblings==0

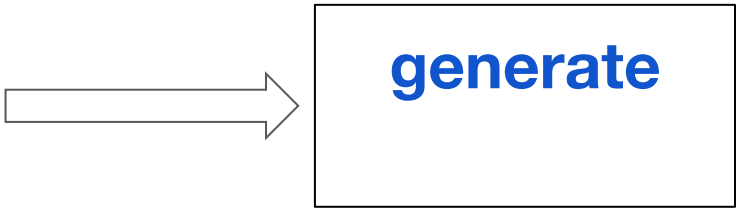
	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	1
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

generate



`generate onlychild`

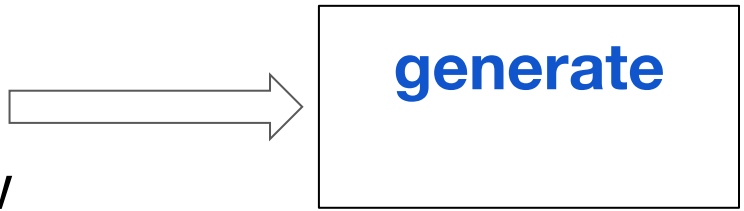
New variable name



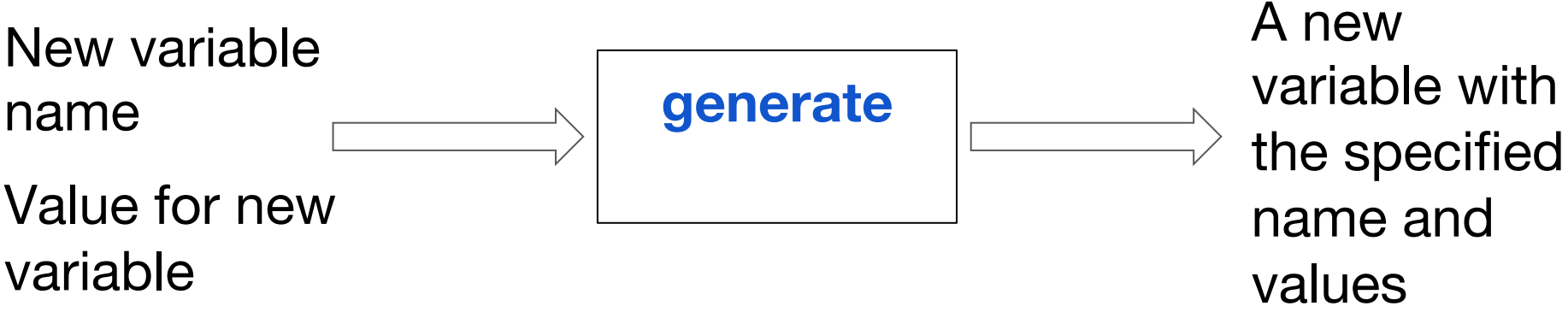
```
generate onlychild = 0
```

New variable name

Value for new variable



```
generate onlychild = 0
```



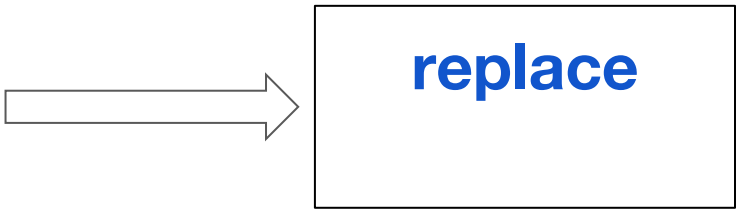
generate onlychild = 0

	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	0
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

replace

`replace onlychild`

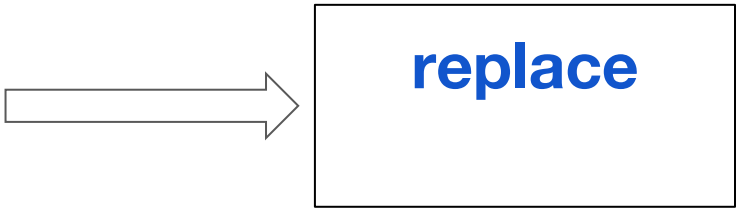
Variable we want to change



```
replace onlychild = 1
```

Variable we want to change

New value

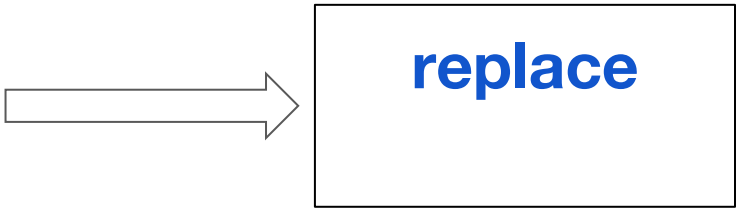


```
replace onlychild = 1 if siblings==0
```

Variable we want to change

New value

If statement specifying a subset



```
replace onlychild = 1 if siblings==0
```

Variable we want to change

New value

If statement specifying a subset

The specified subset of that variable will take on the new value

replace onlychild = 1 if siblings==0

	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	1
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

`replace onlychild = 1 if siblings==0`

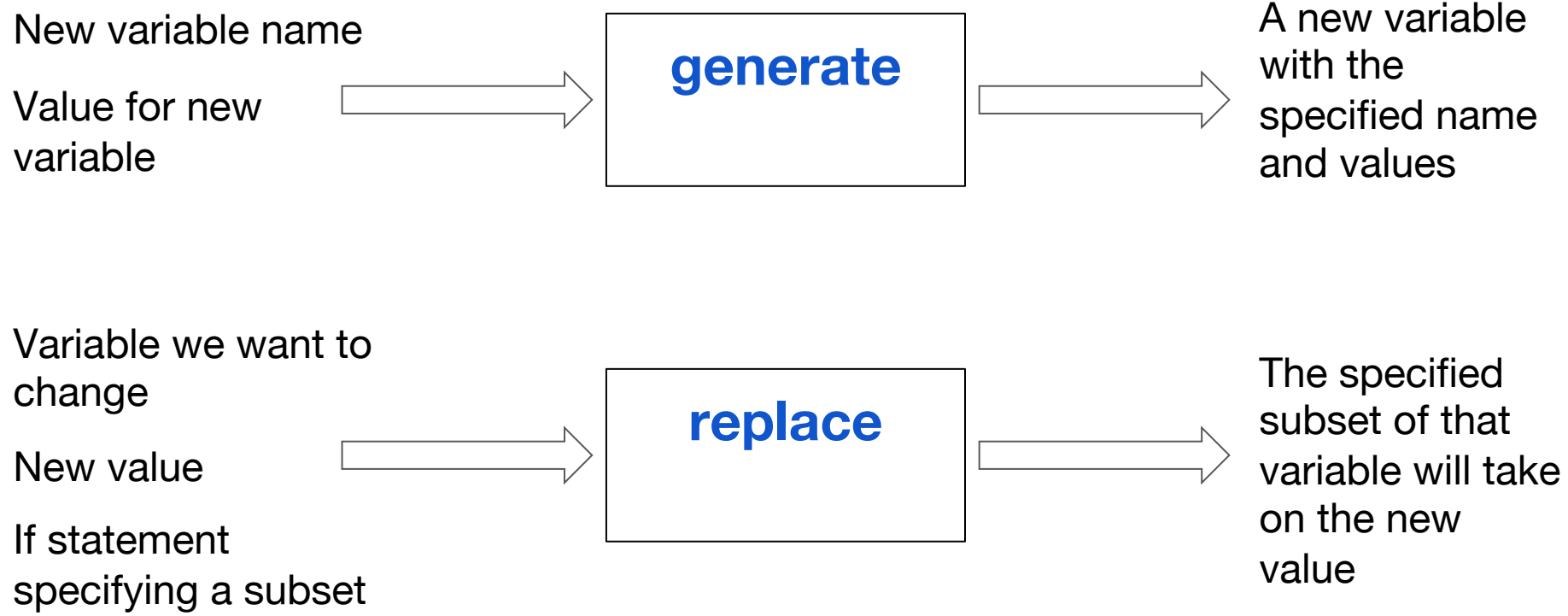
	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	1
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	0
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

But what about this person? How do we deal with **missing data**?

replace onlychild = . if siblings==.

	major	year_school	regions	siblings	height	temp	F_C	cheese	onlychild
1	Spanish	Junior	Northeast,West	1	66	76	F	Brie	0
2	Math	Sophomore	Midwest,West	1	64	63	F	Parmesean	0
3	Sociology	Grad student	Northeast,Midwest,West	3	69	60	F	Gouda	0
4	Sociology	Grad student	Northeast,Midwest,West	1	65	75	F	blue	0
5	Sociology	Grad student	Northeast,West	4	65	75	F	Sharp cheddar	0
6	Sociology	Grad student	Northeast,West	2	83	78	F	Cheddar!!	0
7		Co-term		0	77	0	C	Gouda	1
8		Sophomore	South	1	88	.	.		0
9	Sociology of Education	Grad student	Northeast,West	1	63	80	F	goat	0
10	Undeclared	Freshman	Midwest	.	38	72	F	Sharp cheddar	.
11	Sociology!	Grad student	West	1	68	65	F	a nice sharp gouda	0
12	Sociology	Grad student	Midwest,West	1	70	24	C	feta	0
13	Sociology of Education	Grad student	Northeast,Midwest,West	3	66	75	F	daiya	0

ACTIVITY: “Generate and replace” #14-18



Part 3: Self-Directed Worksheet