DEALING WITH THE MISSING DATA

Missing data affects power, precision, & can cause bias

Case	Age	Gender	Home	Education	Occupation
1	.	Female	No	16	Non-professional
2	22	Male	No	.	Non-professional
3	39	Male	.	20	Professional
4	.	Female	Yes	.	Professional
5	40	.	Yes	16	Non-professional
6	22	Female	No	16	.
7	35	Male	Yes	18	Professional
8	39	Male	Yes	20	Professional

Missing data affects power, precision, & can cause bias

Case	Age	Gender
1	.	Female
2	.	Male
3	39	Male
4	.	Female
5	42	Male
6	.	Female
7	37	Male
8	39	Male

Case	Age	Gender
1	21	Female
2	22	Male
3	39	Male
4	20	Female
5	42	Male
6	18	Female
7	37	Male
8	39	Male

Step 1: Explore missing data

1. How much data is missing?
2. Analyze mechanism of missingness
   - Missing completely at random
   - Missing at random
   - Not missing at random
3. Analyze for patterns of missingness
   - Monotone (systematic)
   - Arbitrary (random)

Missing Completely at Random (MCAR)

Example - Exit Polls

voter	religion	age	gender	party
✗	(row with missing data)			?
✓	complete row			
✓	complete row			
✓	complete row			
✗	(row with missing data)			?
✓	complete row			
✓	complete row			
✓	complete row			
✓	complete row			
✓	complete row			
✓	complete row			
✓	complete row			
✗	(row with missing data)			?

Neither the variable "party" nor rest of the variables are responsible for the missing values

Missing Completely at Random (MCAR)

Example - Exit Polls

voter	religion	age	gender	party
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			

Compare the complete and incomplete rows

data in the "party" column would seemingly cover all the possible parties. No particular party would look suspiciously missing.

voter	religion	age	gender	party
✗ (row with missing party data)				?
✗ (row with missing party data)				?
✗ (row with missing party data)				?

entries in the remaining columns would also look like they are covering all possible religions, age and gender the way they should

Missing at Random (MAR)

Example - Exit Polls

voter	religion	age	gender	party
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			
	✓ complete row			

Compare the complete and incomplete rows

data in the "party" column would seemingly cover all the possible parties. No particular party would look suspiciously missing.

voter	religion	age	gender	party
✗ (row with missing party data)				?
✗ (row with missing party data)				?
✗ (row with missing party data)				?

entries in remaining columns might look a bit odd and you might see missing data from specific religion/age/gender

Missing monotone				Missing arbitrarily			
v1	v2	v3	v4	v1	v2	v3	v4
X	X	X	X	X	X	.	X
X	X	X	X	.	X	X	.
X	X	X	.	X	.	X	.
X	X	.	.	X	X	.	.
X	.	.	.	.	X	X	X

Step 2: Use simple or more advanced methods for handling missing data.

Possible approaches

- remove those rows with any missing data
- switch the missing data with suitable replacements

Likely situations and what might you do?

- ✓ 300 rows with 2-3 rows having missing cells - remove rows with missing data
- ✓ 20 rows with 2-3 rows having missing cells - find replacement

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3		44000	HS	N
4	55	78000	MA	Y
5	23		HS	N
6	25	42000		N
7	35	121000	PhD	Y
8	51	45000	BA	
9			MS	N
10	67	54000	MA	Y

Survey Data

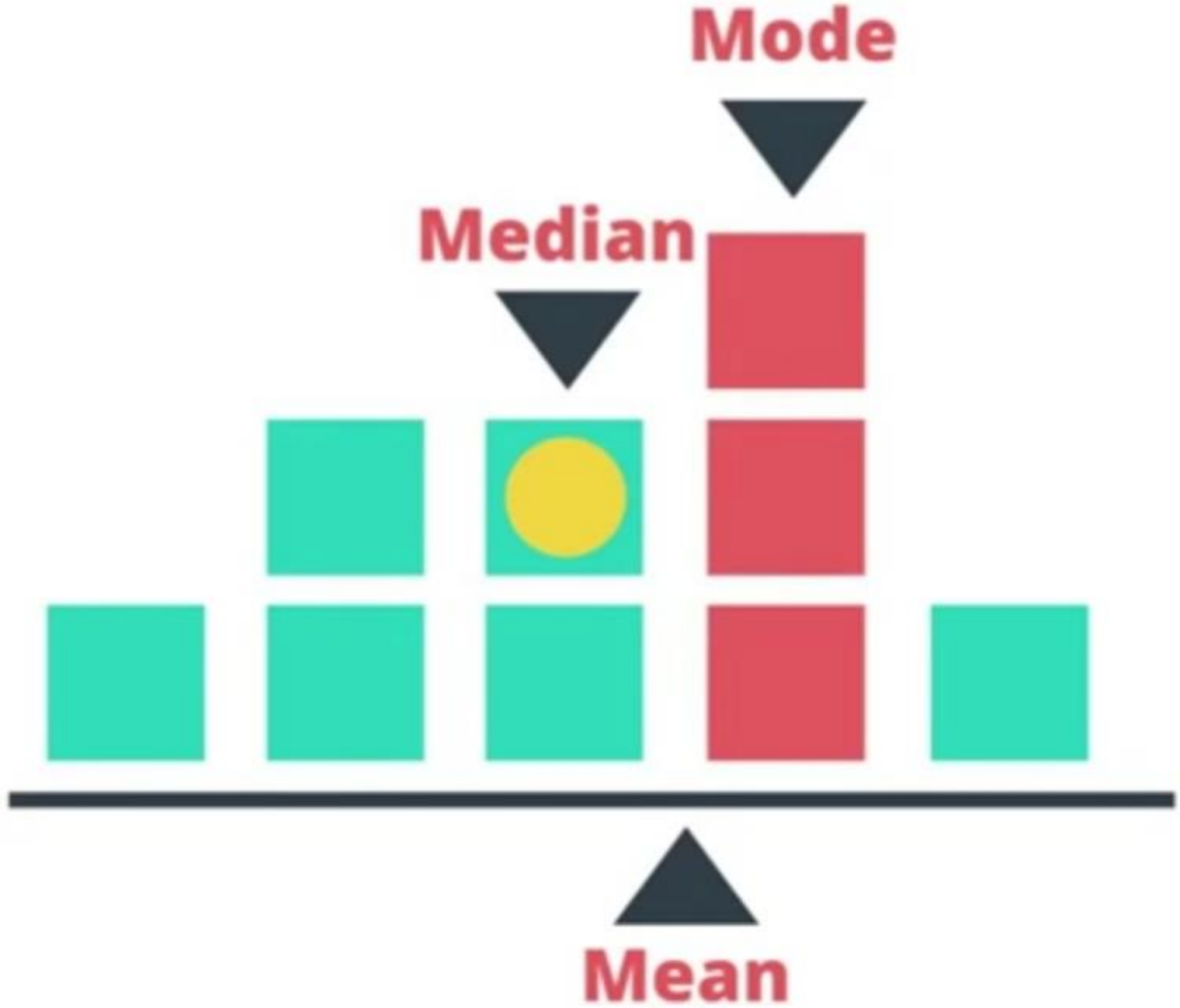
Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
4	55	78000	MA	Y
7	35	121000	PhD	Y
10	67	54000	MA	Y

Survey Data

Delete

Replace Missing Values

- Mean
- Median
- Mode


Mode
Median
Mean

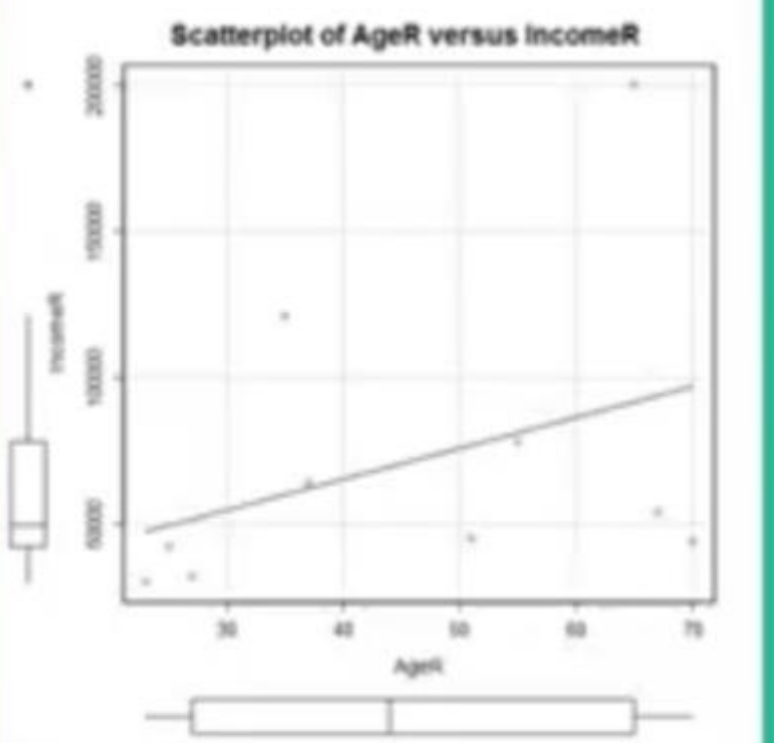
Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	70	44000	HS	N
4	55	78000	MA	Y
5	23	30000	HS	N
6	25	42000	HS	N
7	35	121000	PhD	Y
8	51	45000	BA	N
9	65	200000	MS	N
10	67	54000	MA	Y

Real Data

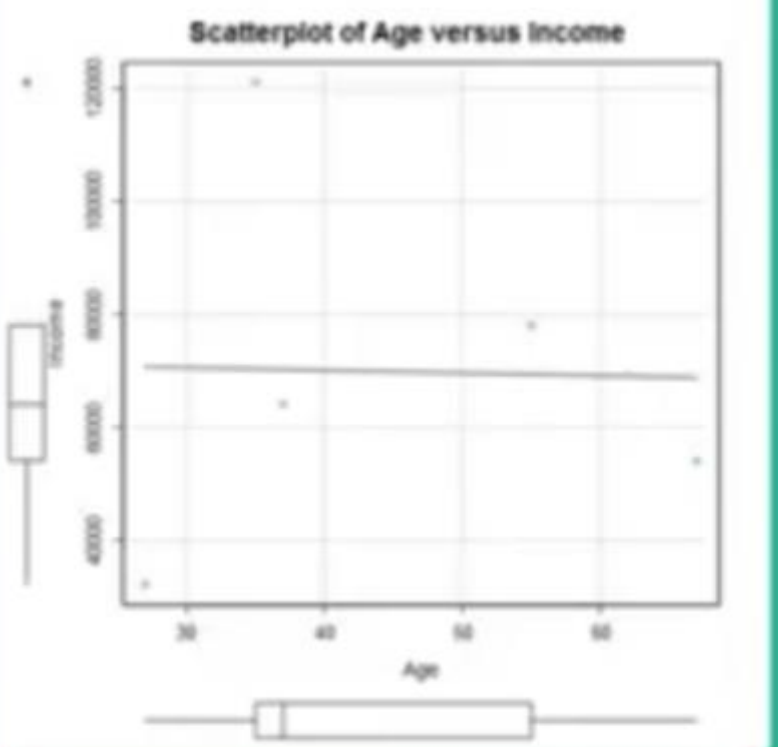
Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	40	44000	HS	N
4	55	78000	MA	Y
5	23	60000	HS	N
6	25	42000	HS	N
7	35	121000	PhD	Y
8	51	45000	BA	N
9	40	60000	MS	N
10	67	54000	MA	Y

Imputed Data (mean)

Real data scatterplot



Deleted data scatterplot



Imputed Scatterplot