Missing data and imputation

INTRODUCTION TO PYTHON IN POWER BI

Jacob H. Marquez
Data Scientist

datacamp

What is missing data?

Common values for "missing":

- null
- NA
- 99
- ""

What is missing data?

Common values for "missing":

- null
- NA
- 99
- ""

entity	year	fished
Australia	1988	153148
Australia	1989	null
Australia	1990	567895
Australia	1991	632987
Australia	1992	643578
Australia	1993	null

Why is data missing?

- A participant forgot or refused to answer a question in a survey
- A participant dropped out of the second part of a study
- There was a glitch in the instrument used to obtain measurements
- Privacy laws restrict the use of data

Is it missing at random?

Missing at random

CITY	Rainfall (inches)			
SEATTLE	2.03	1.13	0.52	4.59
	4.67		2.09	4.53
	0.42	2.60	1.90	
NYC	1.35	3.40	3.75	1.75
		3.93	0.07	3.14
	3.96	3.95		3.60
PARIS	4.72		2.27	2.68
	2.33	2.07	1.06	1.38
		4.29	4.29	1.47

Is it missing at random?

Missing not at random

CITY	Rainfall (inches)			
	4.67	1.75	2.09	4.53
SEATTLE				
	0.42	2.60	1.90	3.14
	1.35	3.40	3.75	1.75
NYC	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
	4.72	4.72	2.27	2.68
PARIS	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

Is it missing at random?

Missing not at random

CITY	Rainfall (inches)			
	4.67	1.75	2.09	4.53
SEATTLE				
	0.42	2.60	1.90	3.14
	1.35	3.40	3.75	1.75
NYC	2.68	3.93	0.07	3.14
	3.96	3.95	0.52	3.60
	4.72	4.72	2.27	2.68
PARIS	2.33	2.07	1.06	1.38
	2.07	4.29	4.29	1.47

- Instrument can't detect low readings
- Certain groups of individuals are unlikely to disclose information

How to address missing data?

Missing not at random

- Pause analysis
- Understand reasons for missing data
- Gather more data
- Clearly document limitations and assumptions made

Missing at random

- Delete the observations
- Add an indicator variable for missing, 1, or not, 0
- Imputation

Imputation

Definition: replacing a missing value with another.

Types of Imputation:

- Mean
- Median
- Mode
- Previous or Next values

Best when 5% of less of the column data is missing.

Remember to sort the values!

Imputation - Example

Missing at random

CITY	Rainfall (inches)			
SEATTLE	2.03	1.13	0.52	4.59
	4.67		2.09	4.53
	0.42	2.60	1.90	
NYC	1.35	3.40	3.75	1.75
		3.93	0.07	3.14
	3.96	3.95		3.60
PARIS	4.72		2.27	2.68
	2.33	2.07	1.06	1.38
		4.29	4.29	1.47

Median imputation

CITY	Rainfall (inches)			
SEATTLE	2.03	1.13	0.52	4.59
	4.67	2.06	2.09	4.53
	0.42	2.60	1.90	2.06
NYC	1.35	3.40	3.75	1.75
	3.50	3.93	0.07	3.14
	3.96	3.95	3.50	3.60
PARIS	4.72	2.30	2.27	2.68
	2.33	2.07	1.06	1.38
	2.30	4.29	4.29	1.47

Dataset

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID
506303	PADS	PADS TO MATCH ALL CUSHIONS	1	4/29/2010 10:43:00 AM	0.001	14249
496725	M	Manual	1	2/3/2010 2:16:00 PM	1.5	13619
502660	M	Manual	6	3/25/2010 5:18:00 PM	1.5	13187
509669	90214S	LETTER "S" BLING KEY RING	10	12/13/2009 3:54:00 PM	1.25	16725

Let's practice!

INTRODUCTION TO PYTHON IN POWER BI

Finding missing data

INTRODUCTION TO PYTHON IN POWER BI

Jacob H. Marquez
Instructor

datacamp

Let's practice!

INTRODUCTION TO PYTHON IN POWER BI