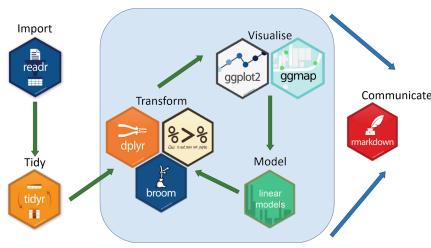


Good Evening!

While you're waiting, please send a browser tab to pollev.com/zahidasghar349 or download the Poll Everywhere app to your phone.



The Taxonomy of Data

Quant Methods in R

Whenever you're learning a new tool, for a long time you're going to suck... But the good news is that is typical, that's something that happens to everyone, and it's only temporary.

-Hadley Wickham

Announcements

1. How to stay in the loop with this class

- atomcamp forum
- google class
- whatsapp group
- send email to TA

2. Getting Help

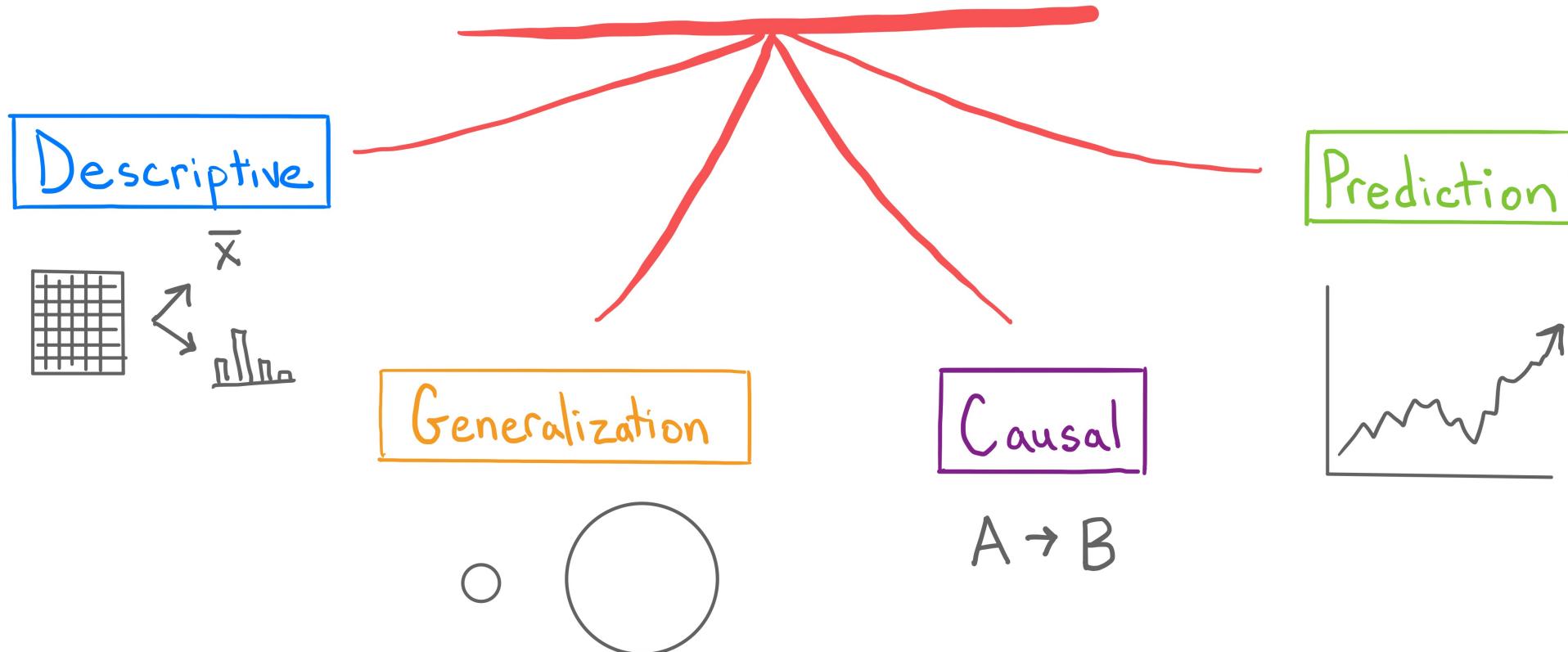
- Send message (or report issue to Simran TA for the course)

Agenda

1. The Taxonomy of Data
2. Sampling Techniques, Response and Explanatory Variables
3. pollev.com/zahidasghar349
4. The Data Frame
5. RStudio.cloud introduction
6. cia, gapminder,penguin and nycbikes data exploration



TO LEARN HOW TO CRITIQUE AND CONSTRUCT CLAIMS MADE USING DATA.



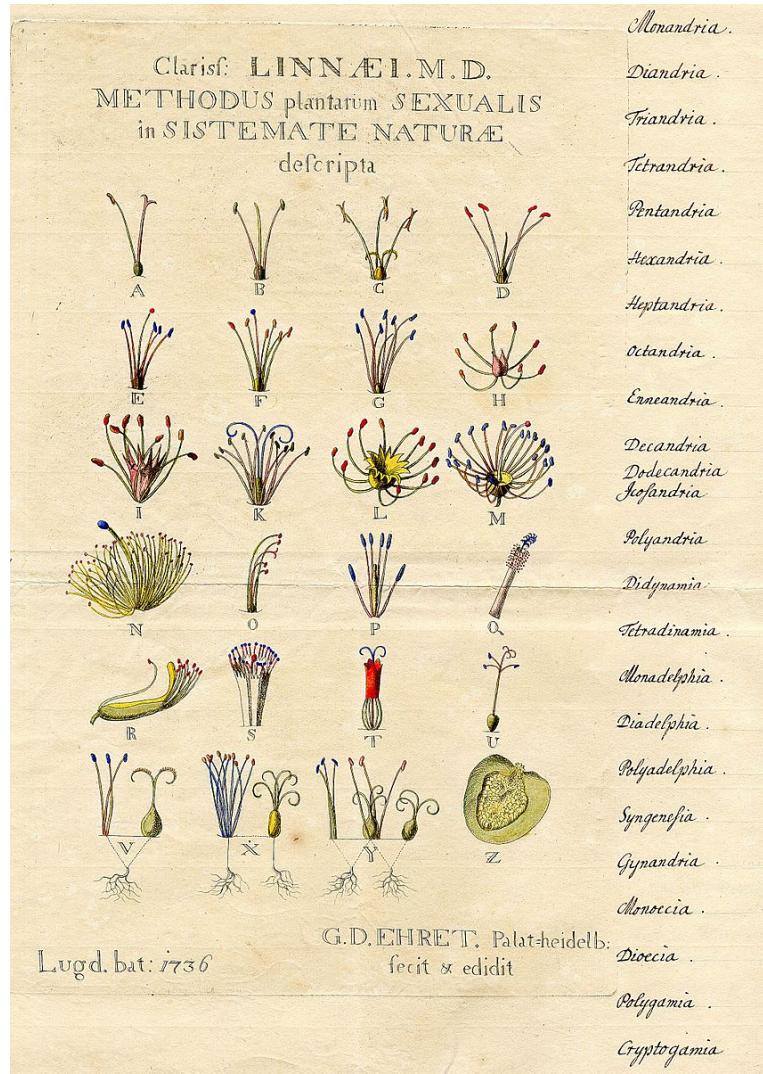
Data defined

Data: (n) An item of (chiefly numerical) information, esp. one obtained by scientific work, a number of which are typically collected together for reference, analysis, or calculation. From latin *datum*: that which is given. Facts.

Variable: A characteristic of an object / unit that can be measured / recorded.

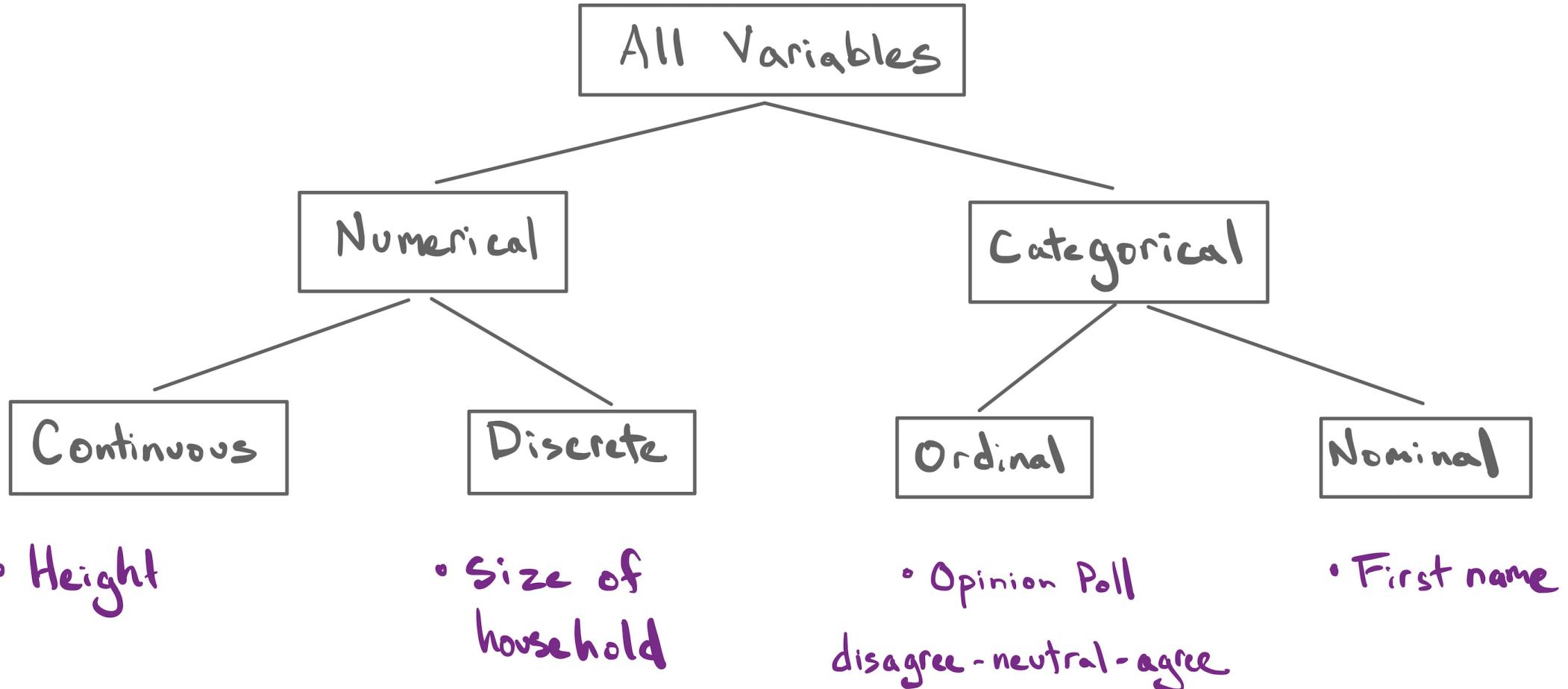
Observations: the collection of objects / units that are being measured / recorded.

Taxonomy



Carl Linnaeus, *Systema Naturae*, 1735.

The Taxonomy of Data



Types of Variables

Numerical: Variables that take numbers as values that are sensible to add, subtract, etc.

Categorical: Variables that take categories as values (called *levels*).

Continuous: A numerical variable that takes values on an interval of the real number line.

Discrete: A numerical variable that takes values that have jumps between them (i.e. an ordered countable set).

Ordinal: A categorical variable with levels that have a natural ordering.

Nominal: A categorical variable with levels with no ordering.

THE TYPE OF VARIABLE INFORMS THE WAY IT SHOULD BE
VISUALIZED AND ANALYZED.

Class volunteers for noting down important questions from chat box?

The Data Frame

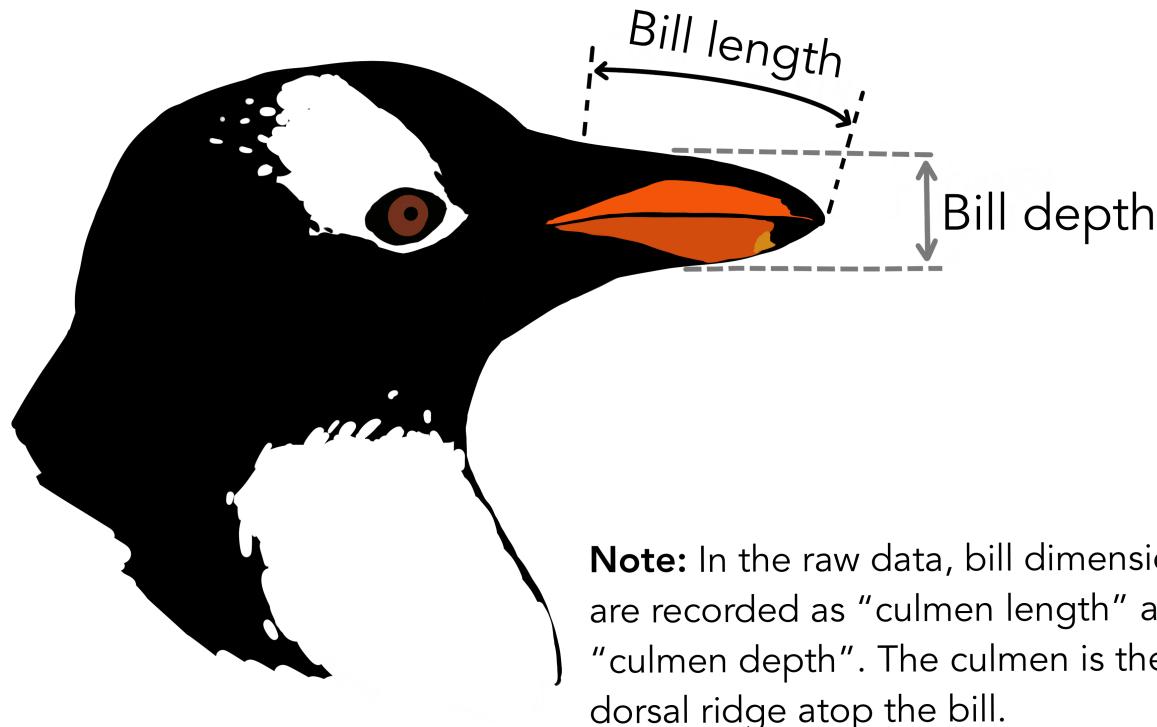
Observing Penguins



Palmer Station, Antarctica

Size measurements for adult foraging penguins near Palmer Station, Antarctica,
K. Gorman 2020.

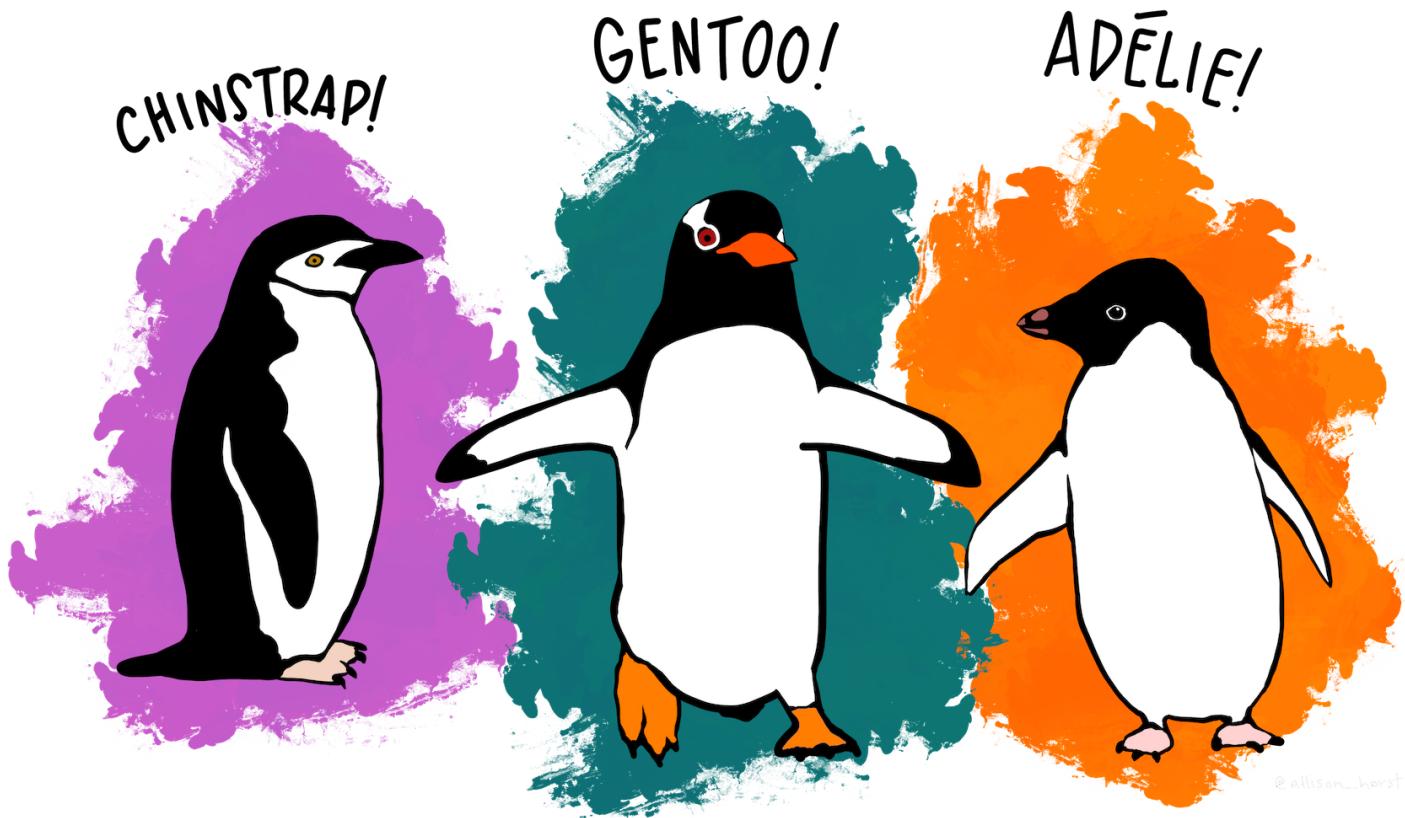
Variables: Beak Length, Beak Depth



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

| These are *continuous* variables.

Variable: Species



| This is a nominal variable.

The Data Frame

bill_depth_mm	bill_length_mm	species
17.9	35.0	Adelie
18.4	50.5	Chinstrap
19.4	41.8	Adelie
17.1	36.4	Adelie

A **data frame** is an array that associates the *observations* (downs the rows) with the *variables* measured on each observation (across the columns). Each cell stores a **value** observed for a variable on an observation.

Unit of observation: The type of the object on which the variables are observed.

| Here: an adult foraging *penguin* near Palmer Station, Antarctica

Non-tidy Data Frames

From Wikipedia, a *contingency table* of sex and handedness.

Sex \ Handedness	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

What is the unit of observation?

What are the variables?

Non-tidy Data Frames

From the World Bank, the number of hospital beds.

Country Name	Country Code	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Ireland	IRL	5.7	5.7	5.5	5.3	5.2	4.9	3.2	3	2.9	2.8
Iran, Islamic Rep.	IRN	1.6	1.6	1.7	1.7	1.7	1.4	1.4	1.7	1.7	0.1
Iraq	IRQ	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3
Iceland	ISL				5.3391	4.2	3.9	3.7	3.6	3.3	3.3
Israel	ISR	3.8	3.7	3.7	3.4	5.8	3.2	3.2	3.2	3.1	3.1
Italy	ITA	4.2	4	4	3.9	3.8	3.7	3.6	3.6	3.5	3.4
Jamaica	JAM	1.39999998	1.8	1.7	2	1.7		1.7	1.8		1.7
Jordan	JOR	1.7	1.7	1.7	1.9	1.9	1.8	1.8	1.8	1.8	1.8
Japan	JPN	14.3000002	13.9	14.1	13.98		13.8	13.7			13.4
Kazakhstan	KAZ	7.7	7.8	7.8	7.8	7.7	7.7	7.6	7.3	7.1	7
Kenya	KEN		1.4		1.4				1.4		
Kyrgyz Republic	KGZ	5.3	5.1	5.1	5.1	5.1	5	4.8	4.8	4.7	4.6
Cambodia	KHM		0.1						0.84	0.7	
Kiribati	KIR		1.5	1.51				1.5	1.4	1.3	
St. Kitts and Nevis	TKN	5.5	6	5.7			5.5	6	4.8	4.8	2.3
Korea, Rep.	KOR	7.0999999	8.6		8.64		12.3	10.3			10.3
Kuwait	KWT	2.2	2.1	1.9	1.9	1.9	1.8	2	2	1.9	2.2
Latin America & Caribbean (e LAC)							1.940519	2.24441397	1.99227147	2.0912725	

What is the unit of observation?

What are the variables?

For next time

1. Read Ch.1 from Openintro
2. Download R and RStudio
3. Practice with RStudio.cloud

