Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

Feb. 24, 2020

5.1: Organizing data in Python

ANNOUNCEMENTS

1. Lab 2 is out, due Wed., Feb. 26, 8p
2. Lab 3 is out Wed., Feb. 26, due Wed., March 4, 8p
3. Homework 2 out today, Mon., Feb. 24, 8p
4. Homework 2 due Mon., March 9, 8p

COURSE JOURNEY: CHECK-IN

Weeks	Topics
1.	Intro. to data science & thinking like a scientist
2.	Thinking like a scientist & causality
3.	Programming preliminaries
4.	Working with data
5.	Organizing data in Python
6.	Visualizations & functions
7.	Midterm exam
…	
The future	Statistics & prediction
	Ethics & data privacy
	Final exam

We are here

KEY TERMS FROM 4.1: WORKING WITH DATA

Population vs. Sample

Random
Large

Two steps in measurement

Conceptualization
Operationalization

"proxy"

Evaluating data

Random errors
Systematic errors
Errors of validity
Errors of exclusion

"orthogonal"

Also coming back: data types!

We will be working with these concepts a lot this week and **forever!**

Outline

1. Tables
2. Finding and importing a dataset
3. Making sense of a dataset

5.2
Organizing a dataset
Descriptive statistics: concepts
Descriptive statistics: code

TABLES

- The fundamental unit of analysis in data science
  - In pandas they are called data frames
  - You'll see abbreviated as "df" in pandas documentation
- Two components: rows and columns
- Rows = observations
  - Each instance of the units we are studying
  - Example: If I'm studying infections rates of a disease, each unit might be a person in the study (i.e., in our sample)
  - And of course: We include people who did and did not become infected! **(Review: why?)**
- Columns = variables
  - Attributes of or pieces of information about, the units we are studying
  - Example: Whether someone became infected with the disease, what temperature their fever was, how old the are

EXAMPLE TABLE: GENERIC

variable1	**variable2**	**variable3**	**variable4**	**variable5**
observation1				
observation1				
observation1				
observation1				

EXAMPLE TABLE: INFECTION RATES

name	infected	age	temp	drink_pump
John Smith				
Bernie Sanders				
Andrea Jones-Rooy				
Liza Minelli				

Dummy, or indicator: Quantitative, or numeric, because it takes on 0 or 1, but represents a qualitative feature

dummy/ bool **string** **float** **dummy/ bool**

"We know no king but the king in the North whose name is **Table**."

I'm fine with that

Data scientists

HOW TO TELL IF YOU'RE WORKING WITH SOMEONE WHO ISN'T A DATA SCIENTIST

Or working with data put together by a non-data scientist

1. Rows and columns are swapped!
2. Dummy variables are backwards
3. Numeric data has symbols in it (, - $ % etc.)
4. Numeric data has units labeled in the data (e.g., "10 pounds" rather than "10")

Outline

1. Tables
2. Finding a dataset
3. Importing a dataset

DATA IN THE REAL WORLD

- Congratulations! You've decided to study something
- You decide to use existing data rather than collect your own
  - (usually a smart and cost-effective first step)
- So you take to the Internet to see what's available
- Here are the four things you will find:
  1. Data in great shape
  2. Data that's usable but needs a bunch of work
  3. Pure crap
  4. A complete lack of data (often a sign that errors of exclusion/invisibility bias is afoot!)

And the one where the data is crap and you're out of time

1. DATA IN GREAT SHAPE

- **First rule of data science**: A lot of data exists in already great shape that was put together and shared by other scientists, experts, and thoughtful people. But somehow, you will almost never encounter that data, especially when you have a deadline
- Great data =
  - Ready-to-use (or ready with minimal changes depending on what you're trying to do)
  - Well-documented with **codebooks** explaining what each variable means and how it's measured (including operationalized & conceptualized)
  - Missing or unusual data is explained or understandable (e.g., "We couldn't find data for Afghanistan in 2002")
    - Explained missing data is great because it helps us think about whether the biases that might arise from it are **random** or **systematic**

EXAMPLE OF (ALMOST) READY-TO-USE DATA

name	price	sex	height	color	location	markings	weight	foaldate	registrations	disciplines	temperament
Captain	5000	Gelding	14.212	Dun	Nantucket, Massachusetts			4-May	Norwegian F	Beginner/Fa	1.005
Eternal Gooc	8500	Gelding	16.205	Chestnut	Brooklyn, Connecticut			3-May	JC - Jockey Cl	Jumper (Com	1.01
Dustys Fly Bc	15000	Gelding	15.192	Grulla	Dallas, Texas		1200 pounds	6-Apr	AQHA - Ame	Beginner/Fa	1.012
A FEDERAL H	8500	Mare	14.999	Grey	HOLSTEIN, Ic	star, strip, & snip. 3 white		5-Apr	AQHA - Ame	Western Plea	1.013
WIMPYS TRA	15000	Gelding	14.999	Palomino	HOWELL, Michigan		1000 pounds	9-Apr	AQHA - Ame	Youth/4-H Hc	1.013
Maximus	2000	Gelding	15.206	Palomino	Weogufka, Alabama			7-Jan		Gaited (Trair	1.014
Times Are Hc	7500	Mare	15.284	Bay	Barneveld, Wisconsin			9-Apr		Western Plea	1.014
C Ya Later Al	9000	Mare	15.186	Sorrel	Camp Spring	White star a	1100 pounds	7-Jan	AQHA - Ame	Hunter (Com	1.019
Ellie	1500	Mare	15.2	Bay	Stoughton, V	Dark bay wit	1500 pounds	May-99		Beginner/Fa	1.02
ONYX	6500	Gelding	87	Black	Buxton, Mair	Solid Black		11-Apr		Beginner/Fa	1.026
savannah	750	Mare		Black	stockton, California			Feb-00		Beginner/Fa	1.027
Elouise GS	5000	Mare	16.11	Chestnut	Markham, Ontario, Canada			1-Sep	ATA - Americ	Jumper Hun	1.034
Bo			13.983	Buckskin	paris, Tennes	buckskin	1200 pounds	Jan-00			1.036
High Dollar			14.3	Bay	Dallas, Texas		1100 pounds	1-Apr	AQHA - Ame	Cutting (Trai	1.036
A-Mill			16.014	Bay	Woodstown, New Jersey			5-May	KWPN - Dutc	Jumper	1.042
Ban			16.188	Chestnut	Queen Anne,	3 Stockings, Blaze		13-Mar	RPSI - Rheinl	Dressage (Pr	1.053
FER		000	Mare	15.213	Pinto	Buxton, Maine			12-May		Jumper Hun	1.055
Awesome	1	00	Mare	5.315	Chestnut	Sanford, Flor	Coronet/Sock		Jan-98	JC - Jockey Cl	Western Rid	1.056

THINGS WE LIKE ABOUT THIS DATASET

As a personal matter of taste, I like the idea of putting put a number on temperament

name	price	sex	height	color	location	markings	weight	foaldate	registrations	disciplines	temperament
Captain	5000	Gelding	14.212	Dun	Nantucket, Massachusetts			4-May	Norwegian F	Beginner/Fa	1.005
Eternal Good	8500	Gelding	16.205	Chestnut	Brooklyn, Connecticut			3-May	JC - Jockey Cl	Jumper (Cor	1.01
Dustys Fly Be	15000	Gelding	15.192	Grulla	Dallas, Texas		1200 pounds	6-Apr	AQHA - Ame	Beginner/Fa	1.012
A FEDERAL H	8500	Mare	14.999	Grey	HOLSTEIN, Ic	star, strip, & snip. 3 white		5-Apr	AQHA - Ame	Western Ple	1.013
WIMPYS TRA	15000	Gelding	14.999	Palomino	HOWELL, Michigan		1000 pounds	9-Apr	AQHA - Ame	Youth/4-H H	1.013
Maximus	2000	Gelding	15.206	Palomino	Weogufka, Alabama			7-Jan		Gaited (Trai	1.014
Times Are H	7500	Mare	15.284	Bay	Barneveld, Wisconsin			9-Apr		Western Ple	1.014
C Ya Later Al	9000	Mare	15.186	Sorrel	Camp Spring	White star a	1100 pounds	7-Jan	AQHA - Ame	Hunter (Com	1.019
Ellie	1500	Mare	15.2	Bay	Stoughton, W	Dark bay wit	1500 pounds	May-99		Beginner/Fa	1.02
ONYX	6500	Gelding	15.287	Black	Buxton, Mair	Solid Black		11-Apr		Beginner/Fa	1.026
savannah	750	Mare	14.005	Black	stockton, California			Feb-00		Beginner/Fa	1.027
Elouise GS	5000	Mare	16.115	Chestnut	Markham, Ontario, Canada			1-Sep	ATA - Ameri	Jumper Hur	1.034
Bo	1000	Gelding	13.983	Buckskin	paris, Tennes	buckskin	1200 pounds	Jan-00			1.036
High Dollar C	6500	Gelding	14.3	Bay	Dallas, Texas		1100 pounds	1-Apr	AQHA - Ame	Cutting (Tra	1.036
A-Millionaire	500	Stallion	16.014	Bay	Woodstown, New Jersey			5-May	KWPN - Dutc	Jumper	1.042
Bandolay	10000	Mare	16.188	Chestnut	Queen Anne,	3 Stockings, Blaze		13-Mar	RPSI - Rheinl	Dressage (P	1.053
FERN	12000	Mare	15.213	Pinto	Buxton, Maine			12-May		Jumper Hur	1.055
Awesome	1500	Mare	15.315	Chestnut	Sanford, Flor	Coronet/Sock		Jan-98	JC - Jockey Cl	Western Ric	1.056

sweet names

within-variable data type consistency

generally clear variable names

THINGS THAT COULD BE BETTER

name	price	sex	height	color	location	markings	weight	foaldate	registrations	disciplines	temperament
Captain	5000	Gelding	14.212	Dun	Nantucket, Massachusetts			4-May	Norwegian F	Beginner/Fa	1.005
Eternal Good	8500	Gelding	16.205	Chestnut	Brooklyn, Connecticut			3-May	JC - Jockey Cl	Jumper (Com	1.01
Dustys Fly Bo	15000	Gelding	15.192	Grulla	Dallas, Texas		1200 pounds	6-Apr	AQHA - Ame	Beginner/Fa	1.012
A FEDERAL H	8500	Mare	14.999	Grey	HOLSTEIN, Ic	star, strip, &	snip, 2 white	5-Apr	AQHA - Ame	Western Plea	1.013
WIMPYS TRA	15000	Gelding	14.999	Palomino	HOWELL, Michigan		1000 pounds	9-Apr	AQHA - Ame	Youth/4-H Ho	1.013
Maximus	2000	Gelding	15.206	Palomino	Weogufka, Alabama			7-Jan		Gaited (Train	1.014
Times Are Ho	7500	Mare	15.284	Bay	Barneveld, Wisconsin			9-Apr		Western Plea	1.014
C Ya Later Al	9000	Mare	15.186	Sorrel	Camp Spring	White star a	1100 pounds	7-Jan	AQHA - Ame	Hunter (Com	1.019
Ellie	1500	Mare	15.2	Bay	Stoughton, W	Dark bay wit	1500 pounds	May-99		Beginner/Fa	1.02
ONYX	6500	Gelding	15.287	Black	Buxton, Mair	Solid Black		11-Apr		Beginner/Fa	1.026
savannah	750	Mare	14.005	Black	stockton, California			Feb-00		Beginner/Fa	1.027
Elouise GS	5000	Mare	16.115	Chestnut	Markham, Ontario, Canada			1-Sep	ATA - Americ	Jumper Hunt	1.034
Bo	1000	Gelding	13.983	Buckskin	paris, Tennes	buckskin	1200 pounds	Jan-00			1.036
High Dollar G	6500	Gelding	14.3	Bay	Dallas, Texas		1100 pounds	1-Apr	AQHA - Ame	Cutting (Trai	1.036
A-Millionaire	500	Stallion	16.014	Bay	Woodstown, New Jersey			5-May	KWPN - Dutc	Jumper	1.042
Bandolay	10000	Mare	16.188	Chestnut	Queen Anne,	3 Stockings, Blaze		13-Mar	RPSI - Rheinl	Dressage (Pr	1.053
FERN	12000	Mare	15.213	Pinto	Buxton, Maine			12-May		Jumper Hunt	1.055
Awesome	1500	Mare	15.315	Chestnut	Sanford, Flor	Coronet/Sock		Jan-93	JC - Jockey Cl	Western Ridi	1.056

unit labeling is inconsistent as well as annoying

missing variables raise questions for me

Horses for Sale

Find horses of all breeds and disciplines for sale across the country!

http://www.equine.com/horses-for-sale

Showing 1-24 of more than 1000 Results, Page 1 of 42

Sort

Default | Sort

Nearby

Postal Code | 100 Miles | Find

Current Location

Breeds

Appaloosa (32)

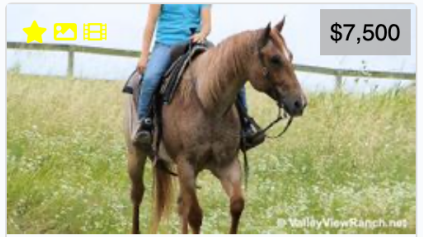
Friesian (31)

Ponies (31)

Paints (All) (27)

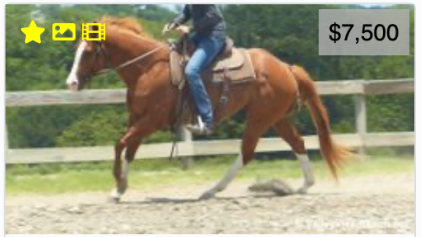
Miniature (25)

This data is only in great shape because someone put it together based on this website



Iota Hear Good News - 15.0HH, 2009', Red Roan AQHA Gelding

Dallas, TX



RJ Skipper - 15.0H, 2007' Sorrel AQHA Gelding

Dallas, TX

Athena Black Cat - 14.2H, 2010', Black AQHA Mare

Dallas, TX

as Cool Cat - 14.3, 2009, l AQHA Gelding

, TX

Holy Smoking Gun - 15.0, 2014' Sorrel APHA Gelding

Dallas, TX

Play Pink - 14.2H, 2009', Cremello AQHA Mare

Dallas, TX

All-Around Level Minded Morgan

Jumper/Eventer/Dressage

Loxley - 14.2, 2012, Red Roan QH Gelding

Blue Smoke Lexus - 2010, 15.2HH, Gray and White Tobiano, Registered Gypsy Vanner Horse Society, Mare

Blue Smoke Lexus - 2010, 15.2HH, Gray and White Tobiano, Registered Gypsy Vanner Horse Society, Mare

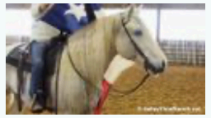
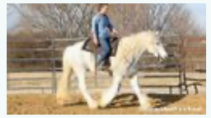
No formal codebook that I am aware of, but we can piece together quite a bit from context

If we really wanted to do a serious study, we could contact the website (or sellers, maybe) for clarity

Photos | Videos





Share: Report Listing Watch Listing

Description

Lexi is stunning!! She is so sweet, quiet, well trained and broke! She is excellent on the trails, in the arena, and in town! She is

Information

Horse Name: Blue Smoke Lexus

Price: $35,000

Location: Dallas, TX

Breed: Gypsy Vanner

Gender: Mare

Height: 15.2 Hands

Color: Grey

Foal Date: January 2010

Markings: Grey and White Tobiano

Disciplines:

- Beginner/Family
- Western Pleasure
- English Pleasure
- Dressage
- Trail Horse

Temperament: 1 (1=Bombproof, 10=Hot)

Contact Seller

ANOTHER EXAMPLE THAT ISN'T RELATED TO HORSES (DIRECTLY)

I am curious about:

Are people better off now than in the past?

I'm starting here because it's well-known source, but there are plenty of others out there

Data source: World Bank

A more thorough study might use multiple sources to minimize the possibility that my results are an artifact of measurement or sampling

Conceptualization: Poverty rates worldwide

Operationalization: Poverty ratios

World Bank Open Data

https://data.worldbank.org

Free and open access to global development data

Search data e.g. GDP, population, Indonesia

Browse by Country or Indicator

MOST RECENT

Many homes in low- and lower middle-income countries lack basic handwashing facilities

H. Kashiwase, Feb 13, 2020

Love, marriage, and development: 4 observations

Daniel Halim, Sergio Rivera, Feb 14, 2020

Energy prices retreated while non-energy prices advanced in January—Pink Sheet

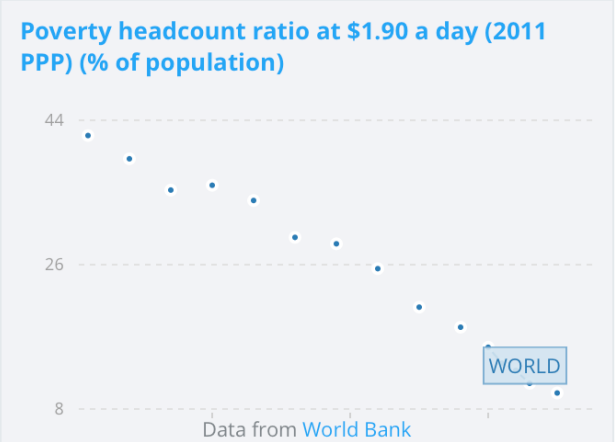
John Baffes, Maria Hazel Macadangdang, Feb 07, 2020

5 steps to reshape economic geography

View all news

View all blogs

WHAT YOU CAN LEARN WITH OPEN DATA



Extreme Poverty

The proportion of the world's population living in extreme poverty has dropped significantly

International Debt Statistics 2020

Oct 02, 2019

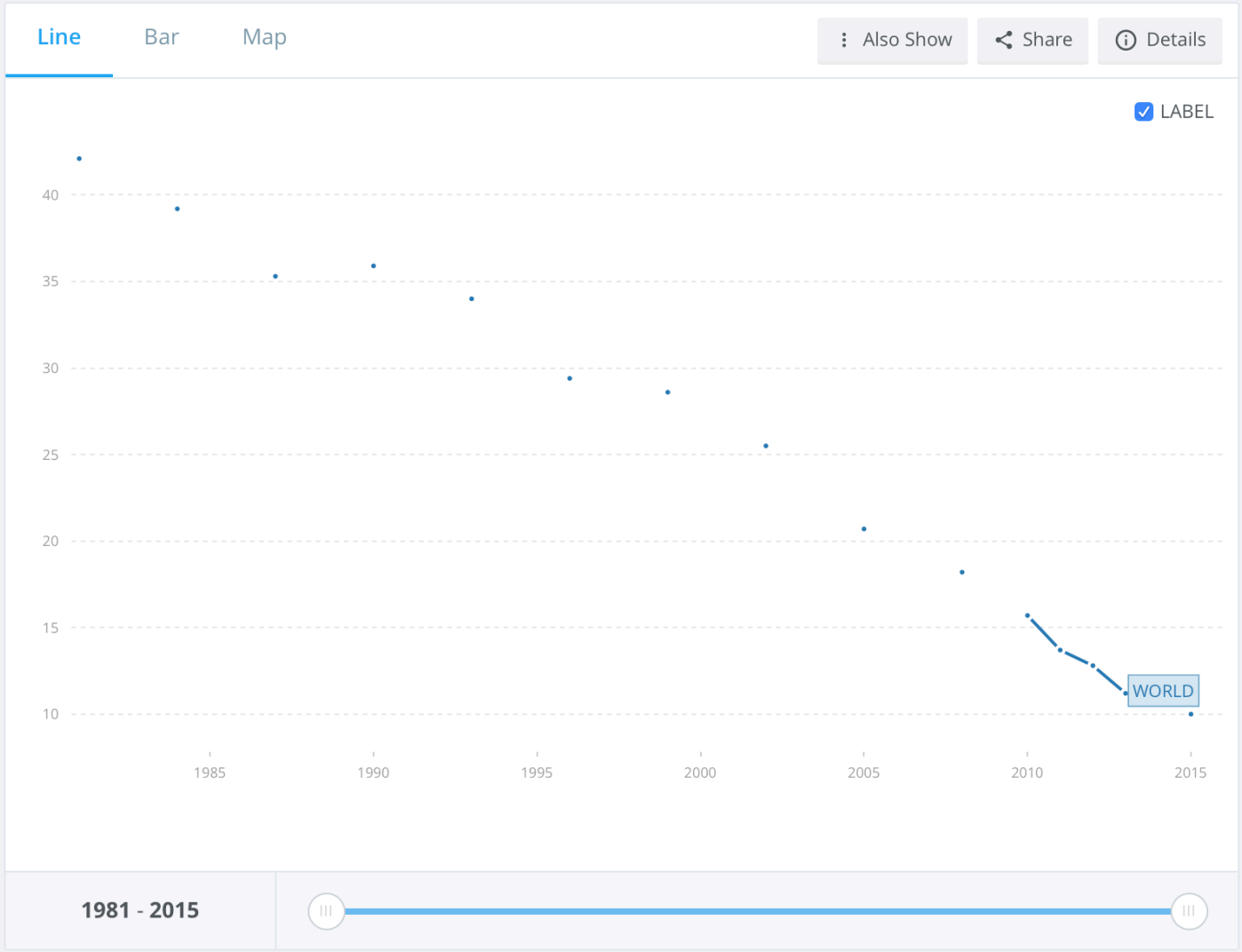
Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population) - World

World Bank, Development Research Group. Data are based on primary household survey data obtained from government statistical agencies and World Bank country departments. Data for high-income economies are from the Luxembourg Income Study database. For more information and methodology, please see PovcalNet (iresearch.worldbank.org/PovcalNet/index.htm).

License : CC BY-4.0

Lots of detailed information about operationalization choices

← This is effectively a codebook



Lots of ways I could have conceptualized

- Poverty headcount ratio at national poverty lines (% of population)
- Poverty gap at $3.20 a day (2011 PPP) (%)
- Rural poverty gap at national poverty lines (%)
- GINI index (World Bank estimate)
- Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)
- Urban poverty gap at national poverty lines (%)
- Rural poverty headcount ratio at national poverty lines (% of rural population)
- Urban poverty headcount ratio at national poverty lines (% of urban population)

Download
CSV XML EXCEL

DataBank
Online tool for visualization and analysis

WDI Tables
Thematic data tables from WDI

For non-data scientists; you don't need this!

THE WORLD BANK
IBRD • IDA

Help us improve this section of the site. Can we get your feedback? Click here ›

I have some feedback.

Sign In

This page is in English Español Français عربي 中文

DataBank | World Development Indicators

Table | Chart | Map | Metadata

Download options

Variables | Layout | Styles | Save | Share | Embed

- Database — Available | Selected 1
- Country — Available | Selected 264
- Series — Available | Selected 1
- Time — Available | Selected 12

Preview

Clear Selection | Add Country (264) Add Series (1) Add Time (12)

Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Afghanistan	..	..	..	..	..	..	..	..	..	..
Albania	..	..	1.1	..	..	..	..	..	..	..
Algeria	..	0.5	..	..	..	..	..	..	..	..
American Samoa	..	..	..	..	..	..	..	..	..	..
Andorra	..	..	..							..
Angola	..	..	..							..
Antigua and Barbuda	..	..	..							..
Argentina	2.2	0.9	0.8							..
Armenia	1.9	2.2	1.6							..
Aruba	..	..	..							..
Australia	0.5	..	..							..
Austria	0.5	0.5	0.5							..
Azerbaijan	..	..	..							..
Bahamas, The	..	..	..							..
Bahrain	..	..	..							..

Source: World Development Indicators. Click on a metadata icon for original sou

Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population) - World

World Bank, Development Research Group. Data are based on primary household survey data obtained from government statistical agencies and World Bank country departments. Data for high-income economies are from the Luxembourg Income Study database. For more information and methodology, please see PovcalNet (iresearch.worldbank.org/PovcalNet/index.htm).

Poverty headcount ratio at national poverty lines (% of population)

Poverty gap at $3.20 a day (2011 PPP) (%)

Rural poverty gap at national poverty lines (%)

GINI index (World Bank estimate)

Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)

Urban poverty gap at national poverty lines (%)

Rural poverty headcount ratio at national poverty lines (% of rural population)

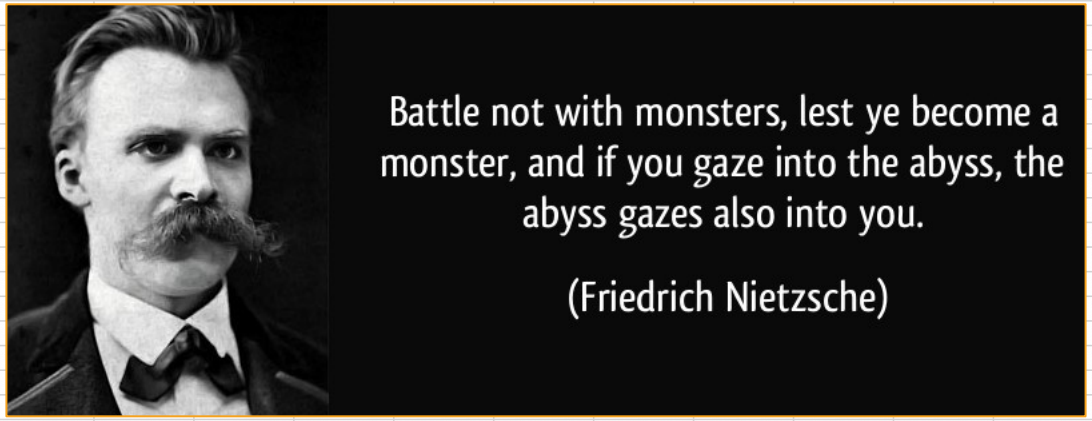
Urban poverty headcount ratio at national poverty lines (% of urban population)

Download
CSV XML EXCEL

DataBank
Online tool for visualization and analysis

WDI Tables
Thematic data tables from WDI

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Data Source	World Development Indicators																							
2																									
3	Last Updated	12/20/19																							
4																									
5	Country Nam	Country Code	Indicator Na	Indicator Coc	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
6	Aruba	ABW	Poverty head	SI.POV.DDAY																					
7	Afghanistan	AFG	Poverty head	SI.POV.DDAY																					
8	Angola	AGO	Poverty head	SI.POV.DDAY																					
9	Albania	ALB	Poverty head	SI.POV.DDAY																					
10	Andorra	AND	Poverty head	SI.POV.DDAY																					
11	Arab World	ARB	Poverty head	SI.POV.DDAY																					
12	United Arab	ARE	Poverty head	SI.POV.DDAY																					
13	Argentina	ARG	Poverty head	SI.POV.DDAY																					0.4
14	Armenia	ARM	Poverty head	SI.POV.DDAY																					
15	American Sa	ASM	Poverty head	SI.POV.DDAY																					
16	Antigua and	ATG	Poverty head	SI.POV.DDAY																					
17	Australia	AUS	Poverty head	SI.POV.DDAY																					
18	Austria	AUT	Poverty head	SI.POV.DDAY																					
19	Azerbaijan	AZE	Poverty head	SI.POV.DDAY																					
20	Burundi	BDI	Poverty head	SI.POV.DDAY																					
21	Belgium	BEL	Poverty head	SI.POV.DDAY																					
22	Benin	BEN	Poverty head	SI.POV.DDAY																					
23	Burkina Faso	BFA	Poverty head	SI.POV.DDAY																					
24	Bangladesh	BGD	Poverty head	SI.POV.DDAY																					
25	Bulgaria	BGR	Poverty head	SI.POV.DDAY																					
26	Bahrain	BHR	Poverty head	SI.POV.DDAY																					
27	Bahamas, Th	BHS	Poverty head	SI.POV.DDAY																					
28	Bosnia and H	BIH	Poverty head	SI.POV.DDAY																					
29	Belarus	BLR	Poverty head	SI.POV.DDAY																					
30	Belize	BLZ	Poverty head	SI.POV.DDAY																					
31	Bermuda	BMU	Poverty head	SI.POV.DDAY																					
32	Bolivia	BOL	Poverty head	SI.POV.DDAY																					
33	Brazil	BRA	Poverty head	SI.POV.DDAY																					
34	Barbados	BRB	Poverty head	SI.POV.DDAY																					
35	Brunei Darus	BRN	Poverty head	SI.POV.DDAY																					
36	Bhutan	BTN	Poverty head	SI.POV.DDAY																					
37	Botswana	BWA	Poverty head	SI.POV.DDAY																					
38	Central Afric	CAF	Poverty head	SI.POV.DDAY																					
39	Canada	CAN	Poverty head	SI.POV.DDAY																					
40	Central Euro	CEB	Poverty head	SI.POV.DDAY																					
41	Switzerland	CHE	Poverty head	SI.POV.DDAY																					
42	Channel Islar	CHI	Poverty head	SI.POV.DDAY																					
43	Chile	CHL	Poverty head	SI.POV.DDAY																					
44	China	CHN	Poverty head	SI.POV.DDAY																					
45	Cote d'Ivoire	CIV	Poverty head	SI.POV.DDAY																					



OK, IF YOU SCROLL TO THE RIGHT THERE IS ACTUALLY DATA

981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
															1.1		
					0	0				1.1	2.1	2.4	2.3	4.1	4.6	4.1	4.6
1				1				0.7						0.7			
														7.3			
											81.1						84.1
													83.1				81.6
		29.9		25.4			37.8			44.2				35.7			
											0						

2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
	30.1										
	0.4				1.1						
2.6	2.6	2.5	2.2	0.9	0.8	0.7	0.7		0.7	0.5	
2.8	1.4	1.9	1.9	2.2	1.6	2.2	2.3	1.9	1.8	1.4	
	0.5		0.5				0.7				
0.5	0.5	0.5	0.5	0.5	0.5	0.2	0.2	0.7			
						71.8					
0.2	0	0	0.2	0.2	0.2	0.2	0	0			
				53.1				49.5			
		55.3					43.7				
			19.6						14.8		
1.7	1.2	1.2	2	2.2	2	1.7	1.5				

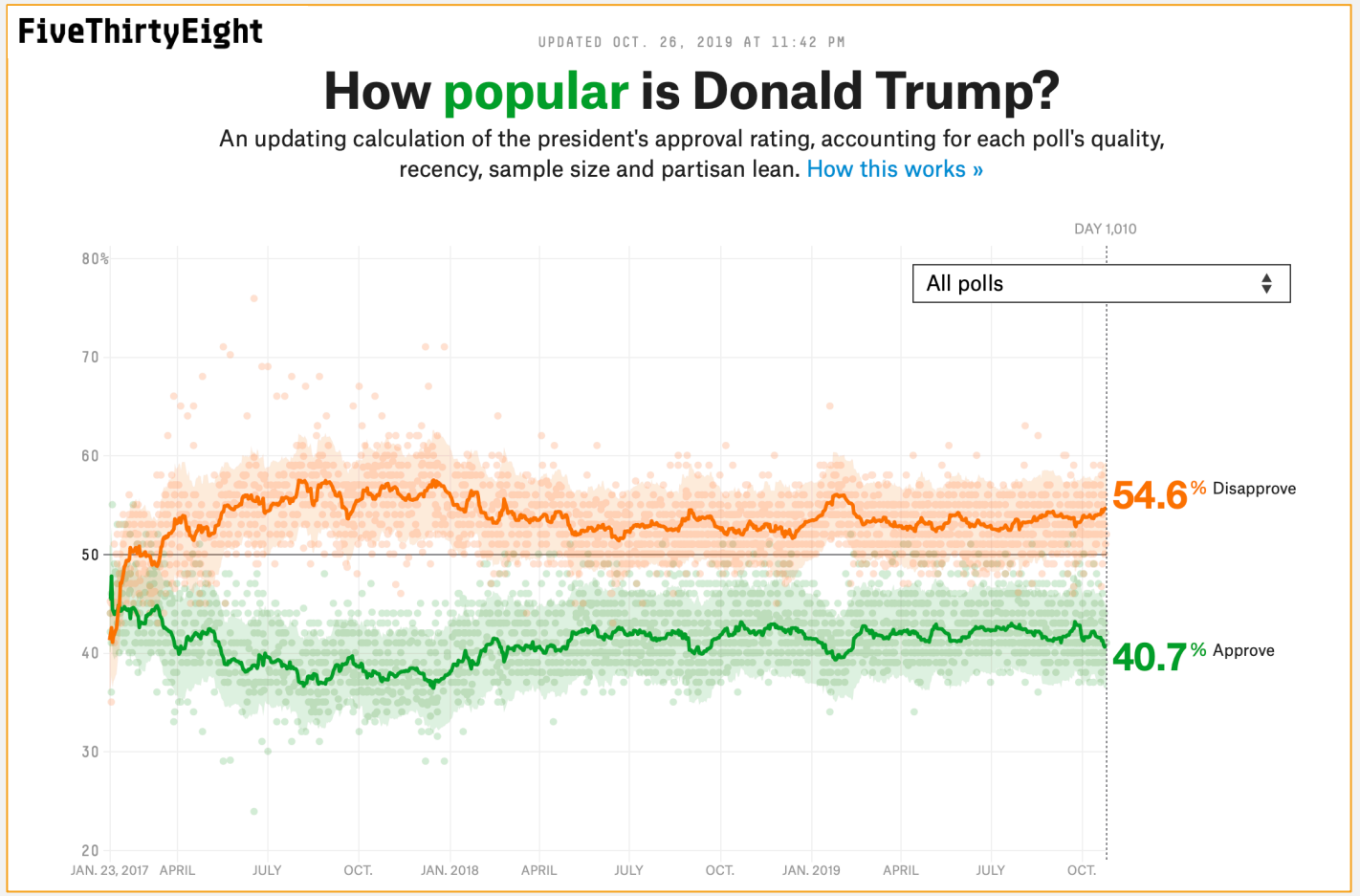
Notice we see more data as we move to more recent years
And in countries that are more developed (loosely defined for now)

These are **systematic** (not random) errors/missing data
We need to keep these in mind if we hope to interpret anything not **orthogonal** to this data

Concepts from 4.1 in action!

Example: If we wanted to say anything about the relationship between poverty ratios and type of government,
We would need to take into account that we have few observations of less-wealthy countries (and adjust inferences accordingly)

2. DATA THAT'S USABLE BUT NEEDS A BUNCH OF WORK



DATES	POLLSTER	GRADE	SAMPLE	WEIGHT	APPROVE	DISAPPROVE	ADJUSTED	
• OCT. 25-26	YouGov	B	1,000 A	0.27	41%	52%	42%	54%
• OCT. 24-25	YouGov	B	1,000 A	0.25	42%	51%	43%	53%
• OCT. 23-24	YouGov	B	1,000 A	0.23	44%	50%	45%	52%
• OCT. 22-24	Rasmussen Reports/Pulse Opinion Research	C+	1,500 LV	0.86	43%	56%	37%	58%
• OCT. 22-24	YouGov	B	1,000 A	0.21	39%	53%	40%	55%
OCT. 21-23	YouGov	B	1,000 A	0.20	40%	53%	41%	55%
OCT. 20-22	YouGov	B	1,500 A	0.28	42%	49%	43%	51%
OCT. 18-22	Ipsos	B+	4,083 A	2.18	39%	55%	40%	54%
OCT. 17-22	McLaughlin & Associates	C-	1,000 LV	0.85	46%	53%	42%	55%
OCT. 8-22	SurveyMonkey	D-	18,101 RV	0.74	44%	54%	41%	53%
OCT. 20-21	YouGov	B	1,000 A	0.19	40%	53%	41%	55%
OCT. 19-21	YouGov	B	1,000 A	0.18	42%	52%	43%	54%
OCT. 18-21	Morning Consult	B-	1,989 RV	0.49	41%	56%	39%	58%
OCT. 18-21	Emerson College	B+	1,000 RV	1.54	43%	47%	42%	49%

Daily Survey: Trump Tweets
October 25 - 26, 2019 - 1000 US adult citizens

YIKES

YouGov

2. President Trump Job Approval

Do you approve or disapprove of the way Donald Trump is handling his job as President?

		Registered voters	Gender		Age (4 category)				Race (4 category)			
	Total	Yes	Male	Female	18-29	30-44	45-64	65+	White	Black	Hispanic	Other
Strongly approve	26%	32%	30%	22%	14%	19%	31%	37%	32%	8%	20%	11%
Somewhat approve	15%	12%	17%	12%	14%	15%	14%	15%	17%	4%	12%	12%
Somewhat disapprove	11%	8%	10%	12%	15%	10%	11%	7%	10%	12%	11%	16%
Strongly disapprove	41%	44%	35%	46%	39%	48%	39%	37%	35%	64%	45%	46%
Not sure	8%	4%	8%	8%	18%	9%	5%	3%	6%	12%	12%	15%
Totals	101%	100%	100%	100%	100%	101%	100%	99%	100%	100%	100%	100%
Unweighted N	(1,000)	(719)	(472)	(528)	(201)	(202)	(396)	(201)	(650)	(141)	(129)	(80)

		Party ID			2016 Vote		Family Income (3 category)			Census Region			
	Total	Dem	Ind	Rep	Clinton	Trump	< $50K	$50-100K	$100K+	Northeast	Midwest	South	West
Strongly approve	26%	6%	20%	65%	4%	68%	21%	33%	28%	22%	30%	27%	22%
Somewhat approve	15%	4%	19%	22%	2%	21%	16%	11%	16%	16%	16%	15%	13%
Somewhat disapprove	11%	13%	12%	6%	7%	3%	12%	11%	10%	13%	10%	9%	12%
Strongly disapprove	41%	73%	36%	3%	83%	4%	42%	42%	42%	44%	35%	39%	46%
Not sure	8%	4%	14%	4%	3%	3%	9%	3%	3%	6%	9%	10%	7%
Totals	101%	100%	101%	100%	99%	99%	100%	100%	99%	101%	100%	100%	100%
Unweighted N	(1,000)	(327)	(421)	(252)	(265)	(283)	(490)	(257)	(130)	(165)	(219)	(353)	(263)

Daily Survey: Trump Tweets
October 25 - 26, 2019 - 1000 US adult citizens

TONS OF POTENTIAL INFORMATION HERE!

YouGov

1. Direction of country
Would you say things in this country today are...

		Registered voters		Gender		Age (4 category)				Race (4 category)			
	Total	Yes		Male	Female	18-29	30-44	45-64	65+	White	Black	Hispanic	Other
Generally headed in the right direction	34%	36%		40%	27%	25%	29%	37%	41%	40%	12%	28%	25%
Off on the wrong track	53%	56%		47%	57%	51%	57%	53%	48%	49%	72%	56%	49%
Not sure	14%	9%		13%	15%	24%	13%	10%	11%	12%	16%	16%	26%
Totals	101%	101%		100%	99%	100%	99%	100%	100%	101%	100%	100%	100%
Unweighted N	(1,000)	(719)		(472)	(528)	(201)	(202)	(396)	(201)	(650)	(141)	(129)	(80)

		Party ID			2016 Vote		Family Income (3 category)			Census Region			
	Total	Dem	Ind	Rep	Clinton	Trump	< $50K	$50-100K	$100K+	Northeast	Midwest	South	West
Generally headed in the right direction	34%	12%	28%	74%	6%	71%	29%	40%	43%	29%	36%	35%	32%
Off on the wrong track	53%	80%	49%	20%	88%	21%	54%	55%	52%	58%	53%	48%	55%
Not sure	14%	7%	24%	7%	6%	8%	17%	4%	5%	13%	11%	17%	14%
Totals	101%	99%	101%	101%	100%	100%	100%	99%	100%	100%	100%	100%	101%
Unweighted N	(1,000)	(327)	(421)	(252)	(265)	(283)	(490)	(257)	(130)	(165)	(219)	(353)	(263)

Daily Survey: Trump Tweets
October 25 - 26, 2019 - 1000 US adult citizens

YouGov

Interviewing Dates	October 25 - 26, 2019
Target population	U.S. citizens, aged 18 and over.
Sampling method	Respondents were selected from YouGov's opt-in Internet panel using sample matching. A random sample (stratified by gender, age, race, education, geographic region, and voter registration) was selected from the 2016 American Community Study. Voter registration was imputed from the November 2016 Current Population Survey Registration and Voting Supplement.
Weighting	The sample was weighted based on gender, age, race, education, and 2016 Presidential vote. The weights range from 0.331 to 2.265, with a mean of one and a standard deviation of 0.323.
Number of respondents	1000
Margin of error	$\pm$ 3.3% (adjusted for weighting)
Survey mode	Web-based interviews
Questions not reported	57 questions not reported.

(LAST PAGE OF MANY PAGES)

TURNING IT INTO NICE DATA!

pollster	media_spons	start_date	end_date	sample_size	sample_type	moe	approve	disapprove	unsure	other	prefer_not_say
YouGov		25-Oct	26-Oct	1000	a	3.3	41	52	8		
YouGov		25-Oct	26-Oct	719	rv	3.3	44	52	4		

Options:

- Manual
- Automate
  - Web scrapers
  - PDF readers (Optical Character Recognition)
- Combination

For future cases

3. DATA THAT IS PURE CRAP

Examples

Grad school admission data on Kaggle

SurveyMonkey service where they provide survey-takers but won't tell you how they were recruited

- Anything without documentation/codebook
- In particular, some minimum need-to-knows:
  - Who/what is in the sample and how were they collected/recruited
  - How each variable is measured (conceptually & operationally)
  - If a lot of missing values, why are they missing?
- All of this should be contained in a codebook
  - Sometimes called a "data dictionary"
  - If there is no such documentation: RUN!
- Of course, "bad" data is also data that was collected sloppily, with instruments that were inaccurate or unreliable, through interview methods that are not well-established, or collected unethically → but you won't know any of this without a codebook or methods section

Good documentation means that in principle, based on the information provided, you could replicate this study

THAT SAID: THERE'S PLENTY OF GOOD DATA THAT IS JUST NOT USEFUL TO YOU

More specifically: what you're trying to understand

- Units/observations aren't in line with what you need:
  - Example: I want to understand income by neighborhood, but all I can find is by county
- Variables are not what you need:
  - Example: I want to understand whether education in STEM or non-STEM is associated with income, but I can only find variables about whether someone has a college degree or not
- Or both:
  - Example: I want to understand factors that contribute to homelessness, but all I can find is locations of homeless shelters
  - Unit I need: Individual homeless people
  - Variables I need: age, family income during childhood, health variables, education, gender, race/ethnicity, LGBTQIA+, veteran status, disability, **or something else I haven't or can't think of!**

Invisibility bias from me!

4. DATA THAT DOESN'T EXIST

- You want to study something, you go looking
- Nothing
- Or not in a version that's useful or relevant to you
- Usually you're dealing with errors of exclusion / invisibility bias
- Something doesn't have data collected for it, or even information about how you could collect it
- Lots of barriers to having data about something:
  - Difficulty/complexity/abstraction: E.g., happiness, freedom, corruption
  - Ethical (and often practical as well): E.g., open adoption, smoking, war
  - Normative/socialized: E.g. Smoking (originally), hormones & anxiety, incarcerated populations, democratization

Outline

1. Tables
2. Finding a dataset
3. Importing a dataset

IMPORTING A DATASET ORDER OF OPERATIONS

- Lots of ways to import, we're using .csv
- Put the .csv file in the working directory, or point to it
- import pandas as pd
- data = pd.read_csv('filename.csv')
- data.head() to inspect that it looks right/get to know what's in there
- And we're off!
- Handy code to get acquainted with a dataset:
  - dtypes, shape, describe, value_counts, mean, sort_values

The Moment You've all been waiting for...

WE'RE GOING TO IMPORT HORSES.CSV!

You can find horses.csv and the example code from Lecture 5.1 on JupyterHub

OFF WE GO

See Lecture 5.1 code on JupyterHub

Lecture 5.1

Importing and and initial exploration of a dataset

In [1]:
```
import pandas as pd                          #we need pandas for this!
data = pd.read_csv('horses.csv')             #this will only work if horses.csv is in the directory where I'm working
```

In [2]:
```
data.head()      #experiment with putting different numbers in the () and see what happens; e.g. data.head(10)
```

Out[2]:

	name	price	sex	height	color	location	markings	weight	foaldate	registrations	disciplines	temperament
0	Captain	5000.0	Gelding	14.212	Dun	Nantucket, Massachusetts	NaN	NaN	4-May	Norwegian Fjord Horse Registry (04-6018-G)	Beginner/Family Cowboy Mounted Shooting Trai...	1.005
1	Eternal Goodness	8500.0	Gelding	16.205	Chestnut	Brooklyn, Connecticut	NaN	NaN	3-May	JC - Jockey Club ()	Jumper (Competed or Shown) Hunter (Competed or...	1.010
2	Dustys Fly Boy	15000.0	Gelding	15.192	Grulla	Dallas, Texas	NaN	1200 pounds	6-Apr	AQHA - American Quarter Horse Association (484...	Beginner/Family (Champion) Youth/4-H Horse (Ch...	1.012
3	A FEDERAL HOLIDAY	8500.0	Mare	14.999	Grey	HOLSTEIN, Iowa	star, strip, & snip. 3 white socks.	NaN	5-Apr	AQHA - American Quarter Horse Association ()	Western Pleasure (Show) (Competed or Shown) Yo...	1.013
4	WIMPYS TRADITIONSTEP	15000.0	Gelding	14.999	Palomino	HOWELL, Michigan	NaN	1000 pounds	9-Apr	AQHA - American Quarter Horse Association (526...	Youth/4-H Horse (Trained) Ranch Horse (Trained...	1.013

In [3]:
```
data.dtypes    #show types of each variable; but how to know this is the code? We tell you, or you SEARCH LIKE MAD
```

Out[3]:
```
name            object
price          float64
sex             object
height         float64
color           object
location        object
markings        object
weight          object
foaldate        object
registrations   object
disciplines     object
```

Outline

1. Tables
2. Finding a dataset
3. Importing a dataset

```
import pandas as pd
data = pd.read_csv('We did it!.csv')
```