http://datahub.ucsd.edu/user-redirect/git-sync?repo=https://github.com/jasongfleischer/CSS202_Pandas_Datavis&subPath=PythonDataScienceHandbook/notebooks/Index.ipynb

```
{
  "lectureTitle": “Data, wrangling, and intuition”,
  "week": 2,
  “date”:  “2022-08-01”
}
```


Jason G. Fleischer, Ph.D.

Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

Many of the slides in this presentation are from material kindly provided by Shannon Ellis

What is data?

Anything that can be stored on a computer

measurements of your height over time, text messages you've sent, Facebook posts your friends have posted, websites you visit, things you buy with a credit card, pictures of your car on speed cameras, audio files downloaded from Spotify, text in emails you've sent to your professors, pictures of your pet, information you fill out in profiles for your school, job, or community organizations, information provided to your physician, list of all the items in your closet, number of clicks on a website's advertisement, list of the sizes of all the shoes ever sold in the world, number of items purchased from a website, information collected in chemistry lab, information entered onto Yelp about your favorite restaurant, photos from your family's vacation.

measurements of your height over time, text messages you've sent, Facebook posts your friends have posted, websites you visit, things you buy with a credit card, pictures of your car on speed cameras, audio files downloaded from Spotify, text in emails you've sent to your professo[illegible]ofiles for your sch[illegible]ovided to your phy[illegible]icks on a website's advertisement, list of the sizes of all the shoes ever sold in the world, number of items purchased from a website, information collected in chemistry lab, information entered onto Yelp about your favorite restaurant, photos from your family's vacation.

If you can enter it into a spreadsheet, take a picture of it, write about it, make a video of it, or record it on audio - then it is probably data.

What is not data?

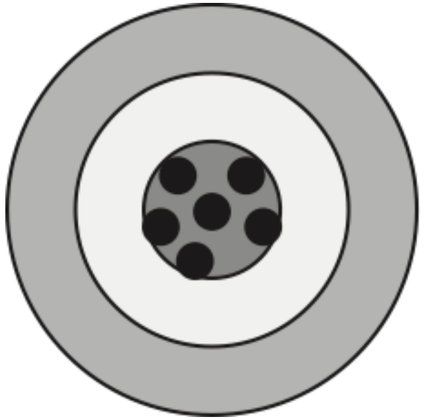
- A subjective experience you have
  - Like an emotion
  - Or a memory
- The real world:
  - Continents, the earth’s mantle, mountains, etc.
  - Trees and other plants, soil teeming with insects and bacteria, etc.

- But if someone measures these things, then THAT is data
- Measurements are NOT the thing/process/idea they measure!!!!!

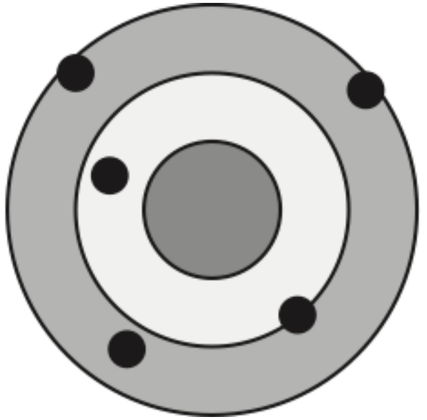
101 Data Problems

(1) Errors of measurement

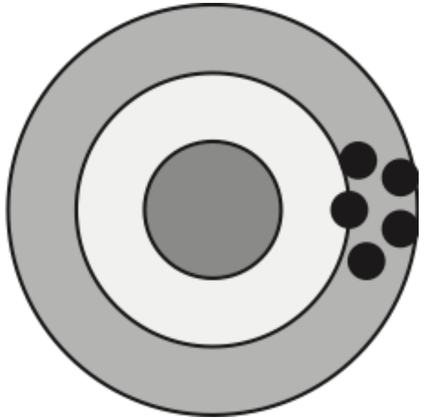
Accurate and precise



Accurate but not precise



Precise but not accurate

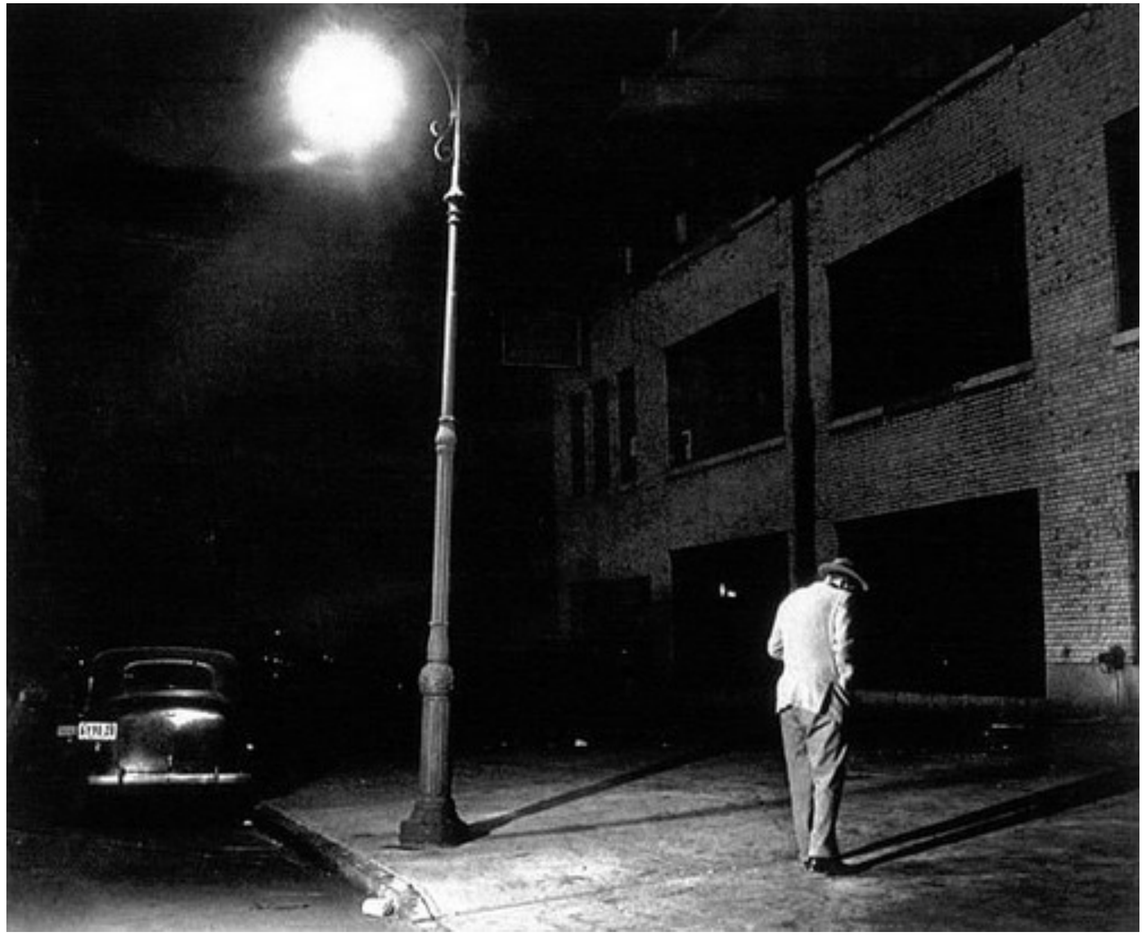


101 Data Problems

(2) Measuring the wrong thing

The Lamppost Problem

aka. Streetlight Effect
aka. The Drunkard's Search



101 Data Problems

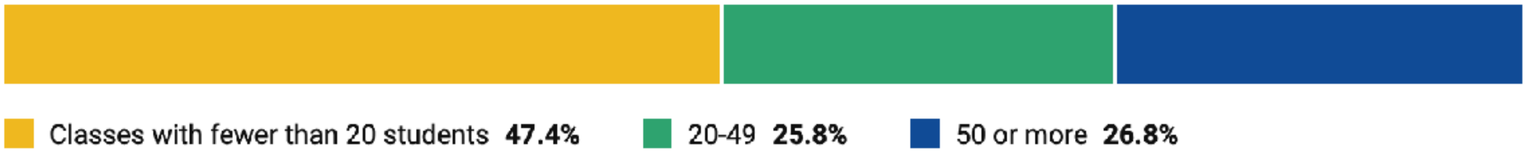
(3) Misunderstanding the data

Survivorship bias

Academic Life at University of California--San Diego

The student-faculty ratio at University of California--San Diego is 19:1, and the school has 47.4% of its classes with fewer than 20 students. The most popular majors at University of California–San Diego include: Biology, General; Mathematics; Economics; International/Global Studies; and Computer Science. The average freshman retention rate, an indicator of student satisfaction, is 94%.

Class Sizes



Classes with fewer than 20 students **47.4%** 20-49 **25.8%** 50 or more **26.8%**

Student-faculty ratio	**19:1**
4-year graduation rate	**65%**

US News & World Report rankings

Becky Yoose
@yo_bj

Hot take - Data is not the new oil. Data is the new glitter:

- Lures humans in with its shininess
- Very easy to accumulate
- Found in places you least likely expect to find it
- Almost impossible to get rid of
- Everyone insists on using it w/o thinking through the consequences

10:24 AM · 10/14/20 · Twitter Web App

1,005 Retweets **90** Quote Tweets

What do we store?
information/values in **variables**

variables == features == attributes

Tabular data

Classic layout

	Variable 1	Variable 2	Variable 3
Observation 1			
Observation 2			
Observation 3			
Observation 4			

Variable Types

Quantitative

- Integers (whole numbers, i.e. 10)
- Continuous: Float or Numeric (200.78)

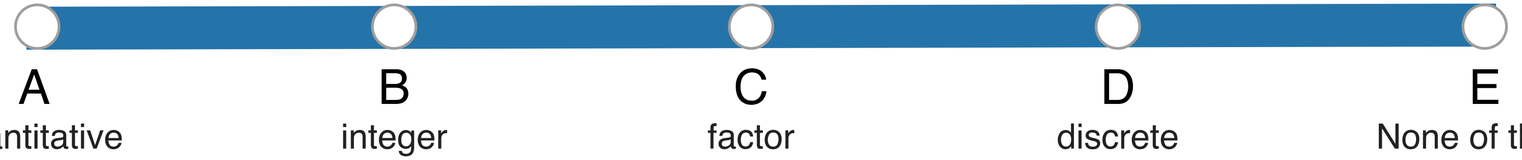
Qualitative

- String or Character (“Hello World”)
- Categorical or Discrete (limited number of options) - (i.e. minor, adult)

Variable Sleuth

I’ve measured your height and entered it into a spreadsheet as 68.2

What type of variable would this be?



Variable Sleuth II

There are 300 species of the iris flower.

What type of variable would `species` be?

A
quantitative

B
categorical

C
continuous

D
float

E
None of these

Data Structures

Structured data

- can be stored in database SQL
- tables with rows and columns
- requires a relational key
- 5-10% of all data

Semi-structured data

- doesn’t reside in a relational database
- has organizational properties (easier to analyze)
- CSV, XML, JSON

Unstructured

- non-tabular data
- 80% of the world’s data
- images, text, audio, videos

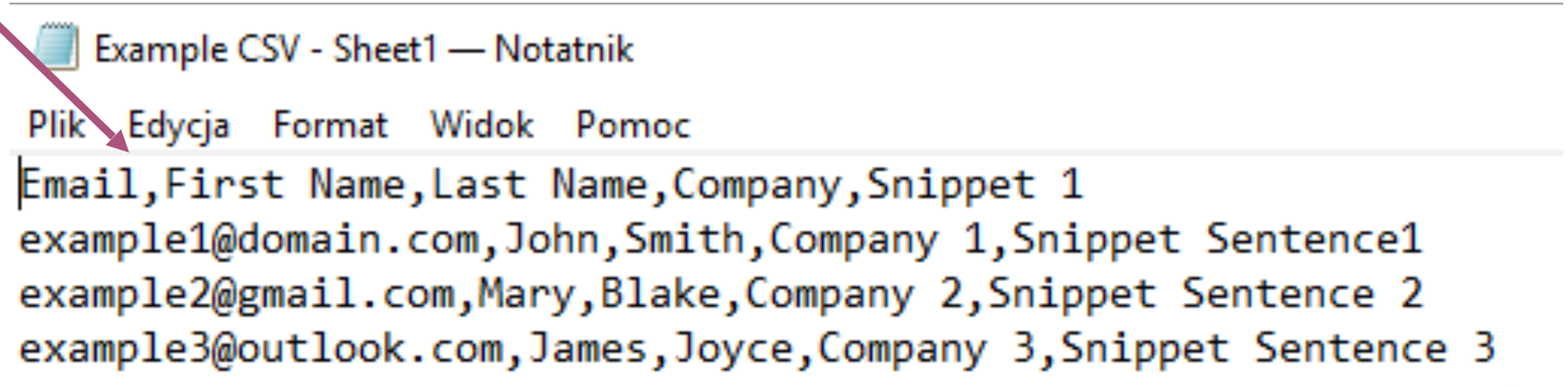
(Semi-)Structured Data

Data that is stored in such a way that it is easy to search and work with. These data are stored in a particular format that adheres to organization principles imposed by the file format. These are the data structures data scientists work with most often.

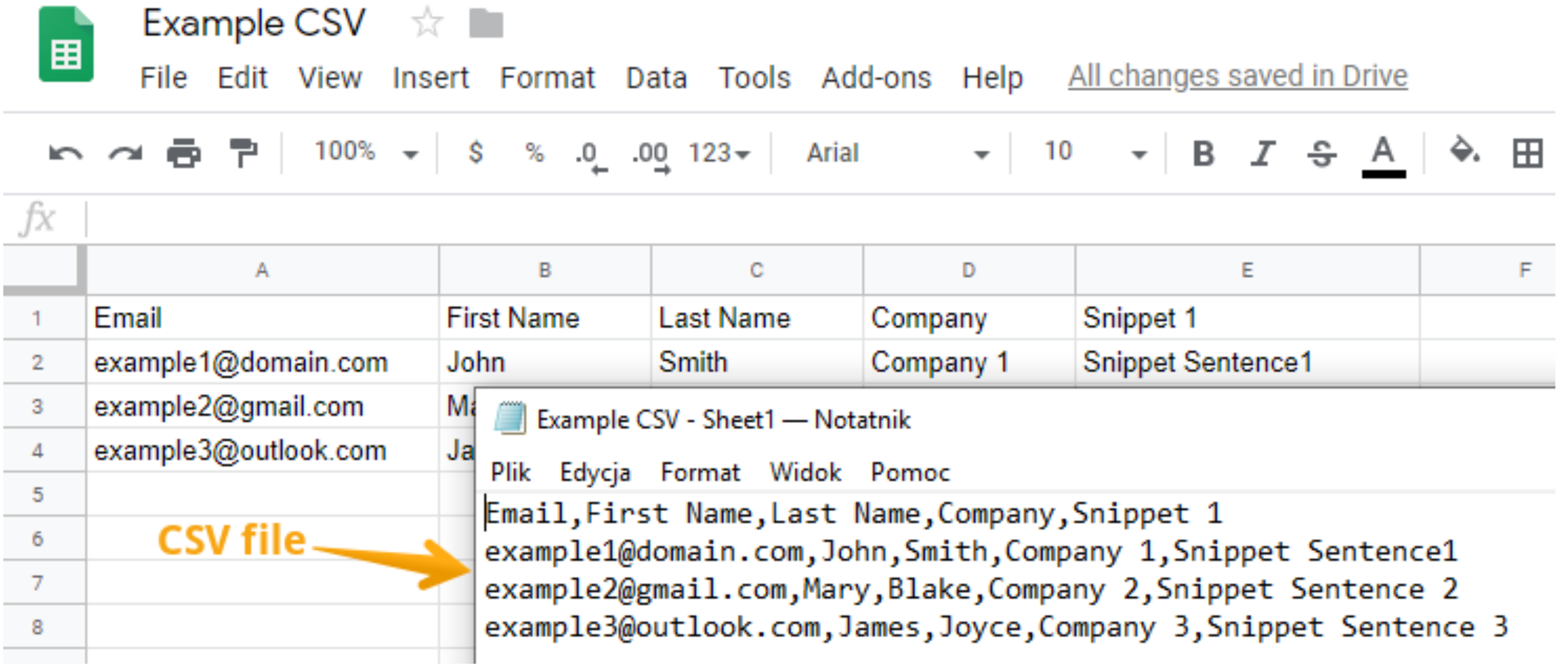
CSVs

Each column separated by a comma

Has the extension ".csv"

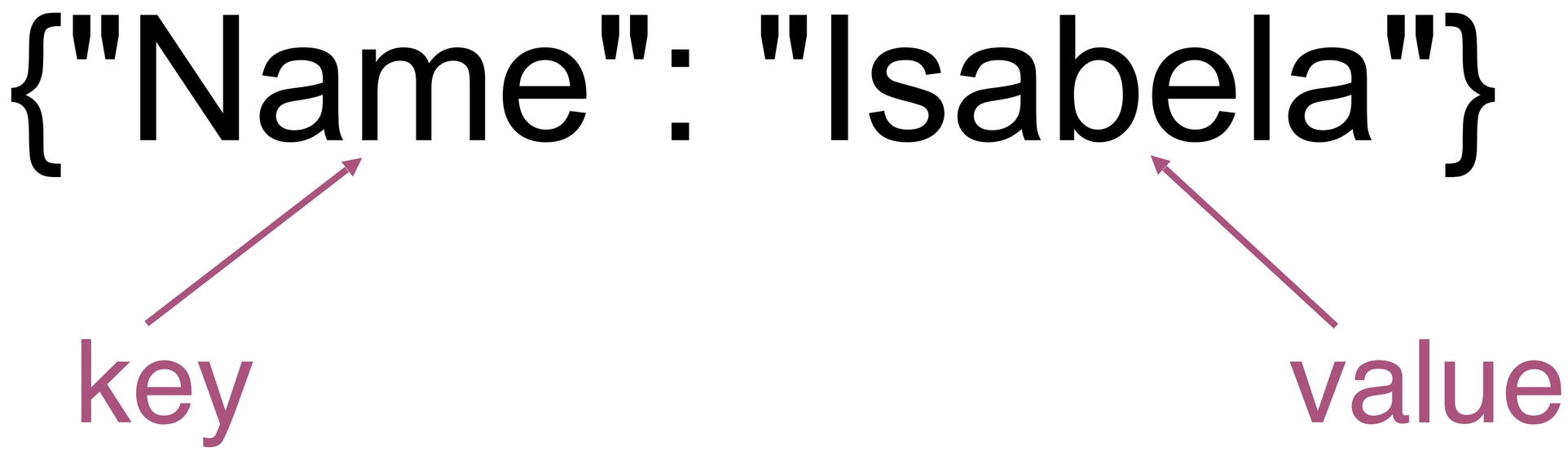


Each row is separated by a new line


Example CSV
File Edit View Insert Format Data Tools Add-ons Help
All changes saved in Drive
100% $ % .0 .00 123 Arial 10
fx
A B C D E F
1 Email First Name Last Name Company Snippet 1
2 example1@domain.com John Smith Company 1 Snippet Sentence1
3 example2@gmail.com
4 example3@outlook.com
5
6
7
8
CSV file
Example CSV - Sheet1 — Notatnik
Plik Edycja Format Widok Pomoc
Email,First Name,Last Name,Company,Snippet 1
example1@domain.com,John,Smith,Company 1,Snippet Sentence1
example2@gmail.com,Mary,Blake,Company 2,Snippet Sentence 2
example3@outlook.com,James,Joyce,Company 3,Snippet Sentence 3

JSON: key-value pairs

nested/hierarchical data

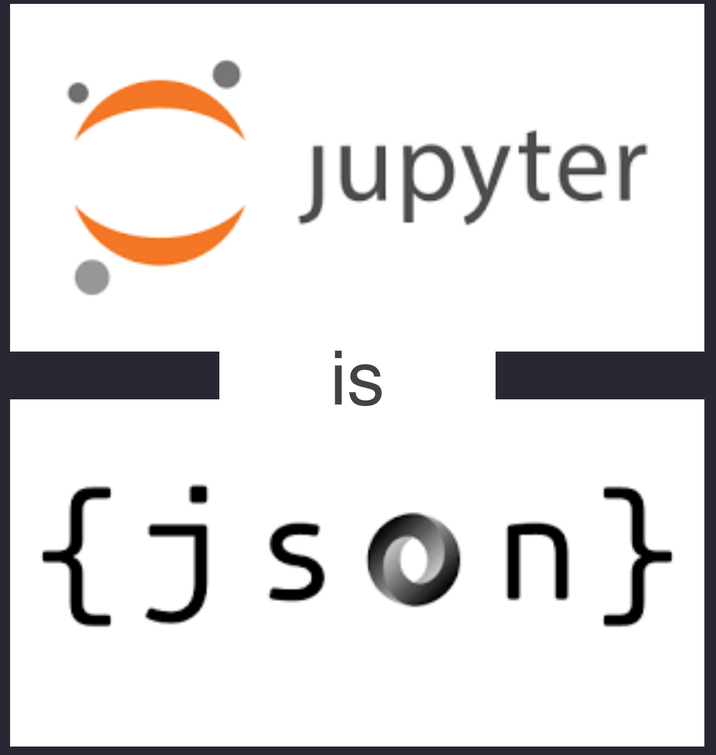


JSON

https://blog.exploratory.io/working-with-json-data-in-very-simple-way-ad7ebcc0bb89

texas_choropleth_example.ipynb | emoji_tsne.ipynb


```
{
 "cells": [
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "This example represents the output the t-SNE dimensionality reduction algorithm on embeddings computed from Unicode emojis using Keras
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "import pandas as pd\n",
    "import holoviews as hv\n",
    "hv.extension('bokeh')"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "## Declaring data"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
```

Jupyter notebooks suck to version control

https://nextjournal.com/schmudde/how-to-version-control-jupyter

```
{
 "cell_type": "code",
 "execution_count": null,
 "metadata": {},
 "outputs": [],
 "source": [
  "import pandas as pd\n",
  "import holoviews as hv\n",
  "hv.extension('bokeh')"
 ]
},
```

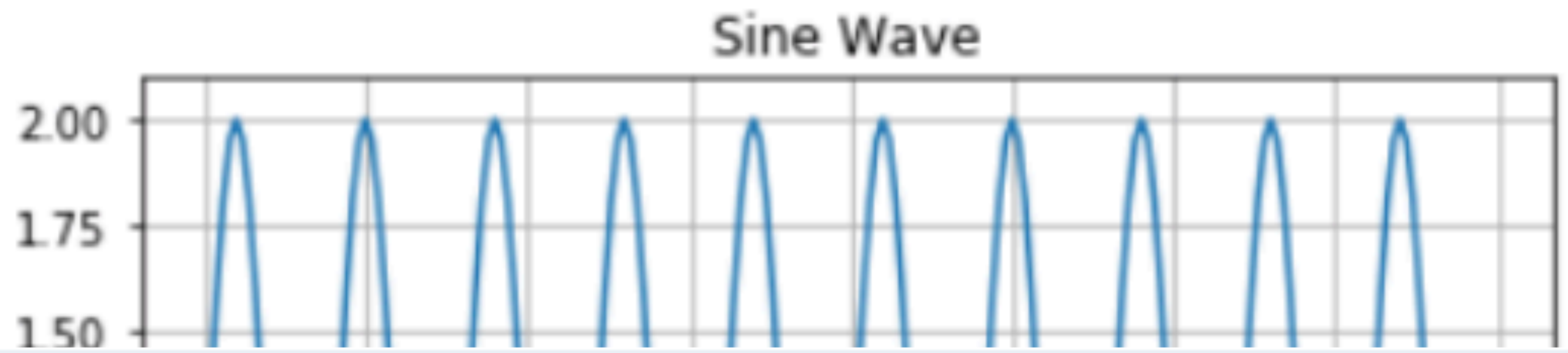
In [10]:
```python
import numpy as np
import matplotlib.pyplot as plt

# Data for plotting
t = np.arange(0.0, 2.0, 0.01)
s = 1 + np.sin((5 * 2)* np.pi * t)

# Note that using plt.subplots below is equivalent to using
# fig = plt.figure() and then ax = fig.add_subplot(111)
fig, ax = plt.subplots()
ax.plot(t, s)

ax.set(xlabel='time (s)', ylabel='voltage (mV)', title='Sine Wave')
ax.grid()
```

Out[10]:



"outputs": [
 {
  "data": {
   "image/png":
"iVBORw0KGgoAAAANSUhEUgAAAYwAAAEWCAYAAAB1xKBvAAAABHNCSVQICAgIfAhkiAAAAAlwSFlzAAALEgAACxIB0t1+/AAAADl
0RVh0U29mdHdhcmUAbWF0cGxvdGxpYiB2ZXJzaW9uIDIuMi4yLCBodHRwOi8vbWF0cGxvdGxpYi5vcmcvhp/UCwAAIABJREFUeJz
svXmcHNd13/s9vc4+2EgABHeQEkVSXGGRFLembFNSPn7Wyy45i5UXh5ZjvcSy4xcr78WK5bwkzvKSeIllOqaVxZKcOJLN+FHc0dx
JEVxAAgQBAiCIdbDP0tPT+80fVdXdmOnl1q17ezBm6HHFFKESNGjBgxYvRDYrkHECNGjBgxVgZigREjRow
YMbQQC4wYMWLEiKGFWGDEiBEjRgwtxAIjRowYMWJoIRYYMWLEiBFDC7HAiBEDEJG/JiKPL/c4QnxxELjBgfGojIXSLyoojMiMg

Jupyter notebooks suck to version control

https://nextjournal.com/schmudde/how-to-version-control-jupyter

Clear Output Manually

The simplest solution is to always clear the output before committing. **Cell → All Output → Clear → Save**. This removes any binary blobs that have been generated by the notebook. There are three main drawbacks:

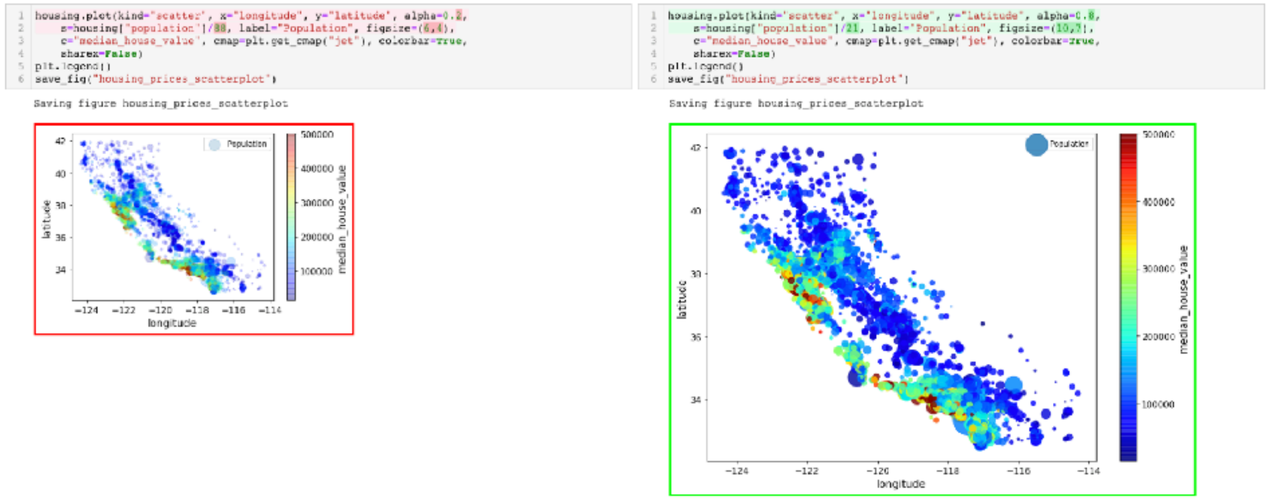
- It is a manual process.
- Collaborators on other machines will need to rerun the notebook to see the output, requiring additional time and setup.

Jupyter notebooks suck to version control

https://nextjournal.com/schmudde/how-to-version-control-jupyter

ReviewNB

ReviewNB is a GitHub app that also offers visual diffing with an interface that looks similar to the traditional Jupyter IDE. Because the outputs are visualized, problems associated with committing binary blobs disappear.

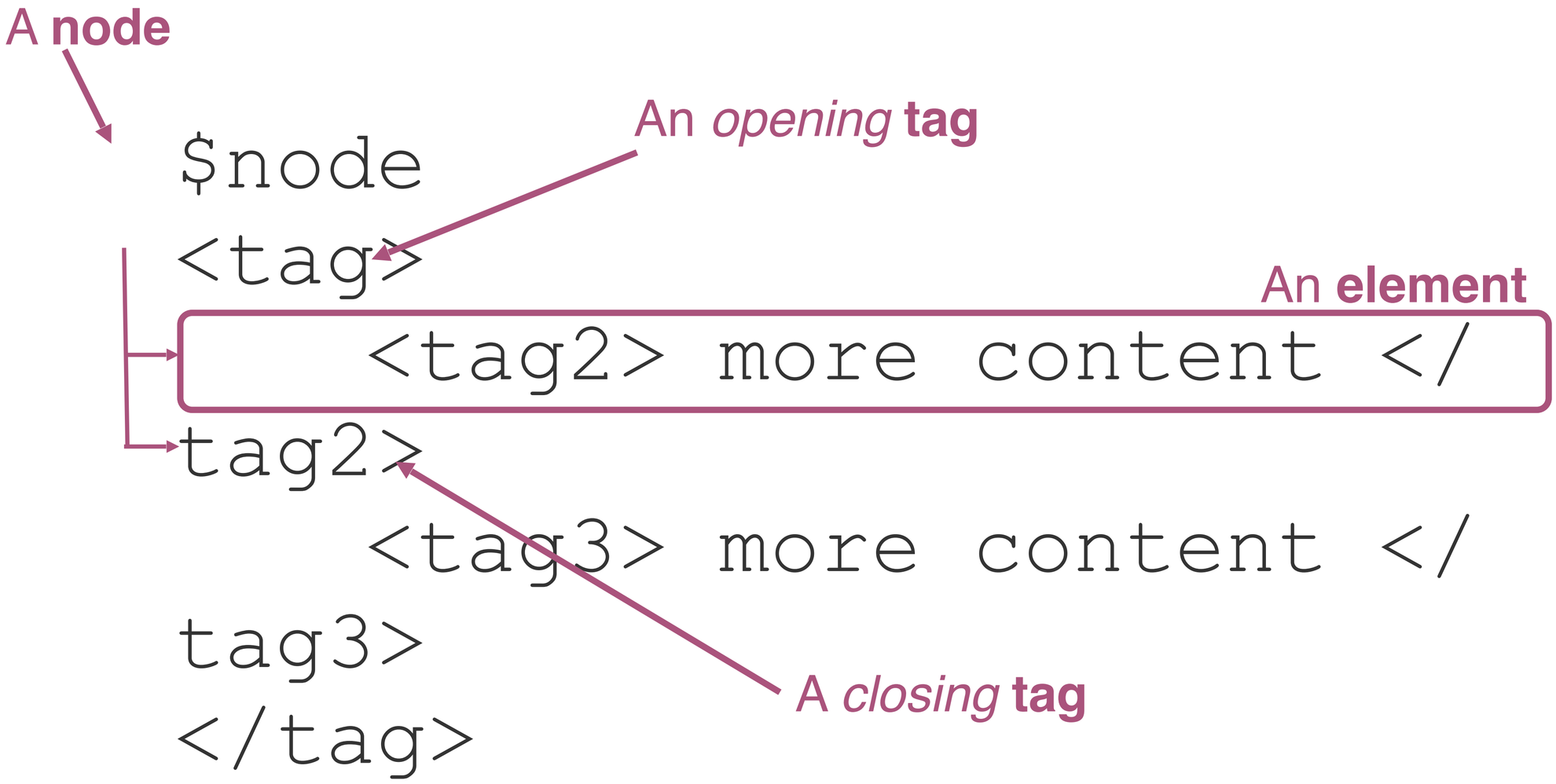


ReviewNB example courtesy of the ReviewNB website

Back to data formats…

Extensible Markup Language (XML): nodes, tags, and elements

nested/hierarchical data



Structured Data
aka databases
[major flavors: SQL, NoSQL]

Relational Databases: A set of interdependent tables

1. Efficient Data Storage
2. Avoid Ambiguity
3. Increase Data Privacy

Information is stored across tables

	unique_identifier	
	AH13JK	
	JJ29JJ	
	CI21AA	

	unique_identifier	
	AH13JK	
	JJ29JJ	
	JJ29JJ	
	XJ11AS	
	CI21AA	

	unique_identifier	
	AH13JK	
	SE92FE	
	CI21AA	

entries are *related* to one another by their unique identifier

restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	**JJ29JJ**	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

health inspections

name	id	inspection_date	inspector	score
Taco Stand	AH13JK	2018-08-21	Sheila	97
Pho Place	**JJ29JJ**	2018-03-12	D'eonte	98
Pho Place	**JJ29JJ**	2018-01-02	Monica	66
Taco Stand	XJ11AS	2018-12-16	Mark	43
Pizza Heaven	CI21AA	2018-08-21	Anh	99

rating

name	id	stars
Taco Stand	AH13JK	4.9
Pho Place	**JJ29JJ**	4.8
Taco Stand	XJ11AS	4.2
Pizza Heaven	CI21AA	4.7

restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	**JJ29JJ**	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

Two different restaurants with the same name

health inspections

name	id	inspection_date	inspector	score
Taco Stand	AH13JK	2018-08-21	Sheila	97
Pho Place	**JJ29JJ**	2018-03-12	D'eonte	98
Pho Place	**JJ29JJ**	2018-01-02	Monica	66
Taco Stand	XJ11AS	2018-12-16	Mark	43
Pizza Heaven	CI21AA	2018-08-21	Anh	99

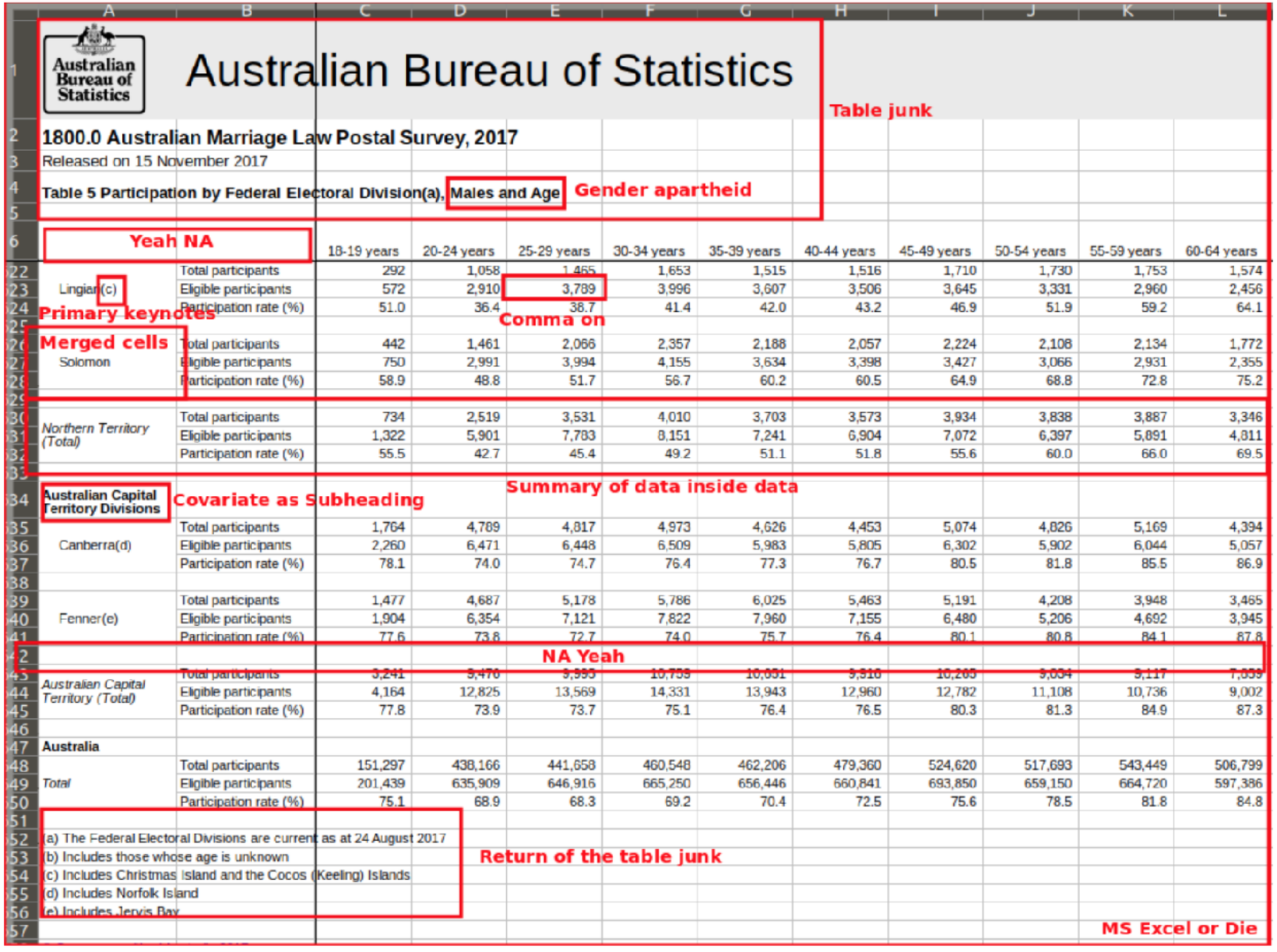
rating

name	id	stars
Taco Stand	AH13JK	4.9
Pho Place	**JJ29JJ**	4.8
Taco Stand	XJ11AS	4.2
Pizza Heaven	CI21AA	4.7

Data wrangling

Image from kidsandhorses.org

"Good data scientists understand, in a deep way, that the heavy lifting of cleanup and preparation isn't something that gets in the way of solving the problem: it is the problem."
- DJ Patil



Australian Bureau of Statistics

Australian Bureau of Statistics

Table junk

1800.0 Australian Marriage Law Postal Survey, 2017

Released on 15 November 2017

Table 5 Participation by Federal Electoral Division(a), Males and Age Gender apartheid

Yeah NA

		18-19 years	20-24 years	25-29 years	30-34 years	35-39 years	40-44 years	45-49 years	50-54 years	55-59 years	60-64 years
Lingiari(c)	Total participants	292	1,058	1,465	1,653	1,515	1,516	1,710	1,730	1,753	1,574
	Eligible participants	572	2,910	3,789	3,996	3,607	3,506	3,645	3,331	2,960	2,456
	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.9	51.9	59.2	64.1
Solomon	Total participants	442	1,461	2,066	2,357	2,188	2,057	2,224	2,108	2,134	1,772
	Eligible participants	750	2,991	3,994	4,155	3,634	3,398	3,427	3,066	2,931	2,355
	Participation rate (%)	58.9	48.8	51.7	56.7	60.2	60.5	64.9	68.8	72.8	75.2
Northern Territory (Total)	Total participants	734	2,519	3,531	4,010	3,703	3,573	3,934	3,838	3,887	3,346
	Eligible participants	1,322	5,901	7,783	8,151	7,241	6,904	7,072	6,397	5,891	4,811
	Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0	69.5
Australian Capital Territory Divisions											
Canberra(d)	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169	4,394
	Eligible participants	2,260	6,471	6,448	6,509	5,983	5,805	6,302	5,902	6,044	5,057
	Participation rate (%)	78.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5	86.9
Fenner(e)	Total participants	1,477	4,687	5,178	5,786	6,025	5,463	5,191	4,208	3,948	3,465
	Eligible participants	1,904	6,354	7,121	7,822	7,960	7,155	6,480	5,206	4,692	3,945
	Participation rate (%)	77.6	73.8	72.7	74.0	75.7	76.4	80.1	80.8	84.1	87.8
Australian Capital Territory (Total)	Total participants	3,241	9,476	9,995	10,759	10,651	9,516	10,265	9,034	9,117	7,859
	Eligible participants	4,164	12,825	13,569	14,331	13,943	12,960	12,782	11,108	10,736	9,002
	Participation rate (%)	77.8	73.9	73.7	75.1	76.4	76.5	80.3	81.3	84.9	87.3
Australia											
Total	Total participants	151,297	438,166	441,658	460,548	462,206	479,360	524,620	517,693	543,449	506,799
	Eligible participants	201,439	635,909	646,916	665,250	656,446	660,841	693,850	659,150	664,720	597,386
	Participation rate (%)	75.1	68.9	68.3	69.2	70.4	72.5	75.6	78.5	81.8	84.8

(a) The Federal Electoral Divisions are current as at 24 August 2017
(b) Includes those whose age is unknown
(c) Includes Christmas Island and the Cocos (Keeling) Islands
(d) Includes Norfolk Island
(e) Includes Jervis Bay

Primary keynotes

Merged cells

Comma on

Summary of data inside data

Covariate as Subheading

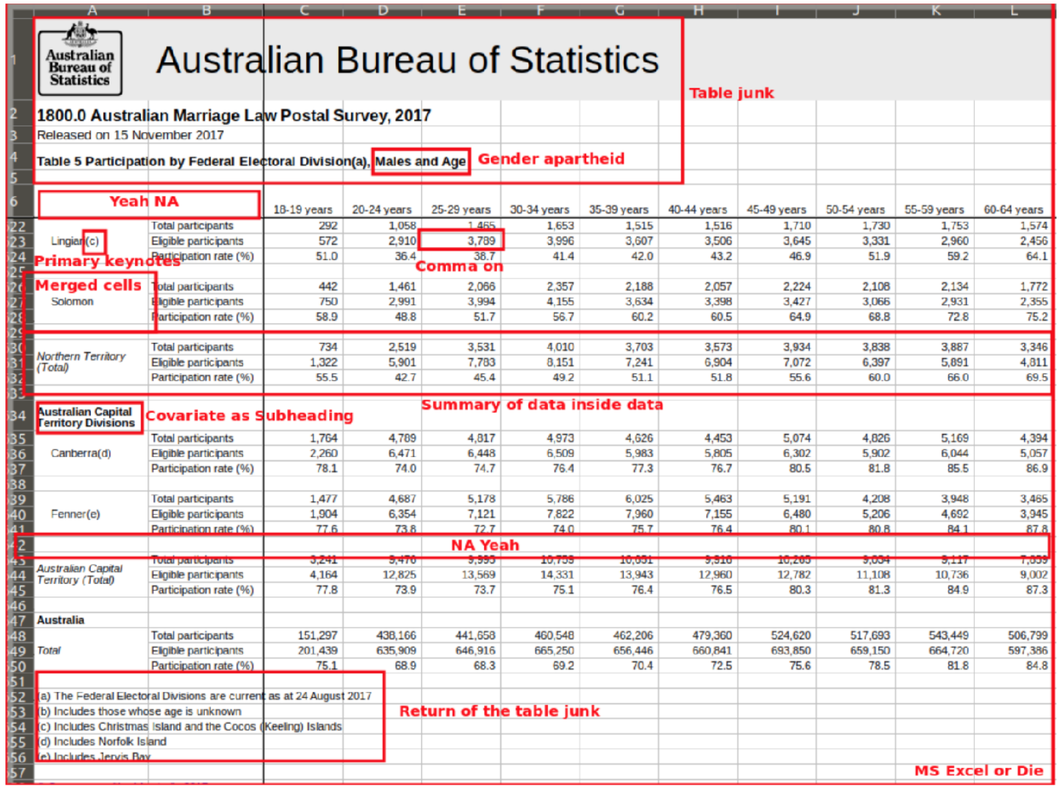
NA Yeah

Return of the table junk

MS Excel or Die

https://medium.com/@miles.mcbain/tidying-the-australian-same-sex-marriage-postal-survey-data-with-r-5d35cea07962

untidy data



Australian Bureau of Statistics

Table junk

1800.0 Australian Marriage Law Postal Survey, 2017

Released on 15 November 2017

Table 5 Participation by Federal Electoral Division(a), Males and Age — Gender apartheid

Yeah NA

		18-19 years	20-24 years	25-29 years	30-34 years	35-39 years	40-44 years	45-49 years	50-54 years	55-59 years	60-64 years
Lingiari(c)	Total participants	292	1,058	1,465	1,653	1,515	1,516	1,710	1,730	1,753	1,574
	Eligible participants	572	2,910	3,789	3,996	3,607	3,506	3,645	3,331	2,960	2,456
Primary keynotes	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.9	51.9	59.2	64.1
Merged cells	Total participants	442	1,461	2,066	2,357	2,188	2,057	2,224	2,108	2,134	1,772
Solomon	Eligible participants	750	2,991	3,994	4,155	3,634	3,398	3,427	3,066	2,931	2,355
	Participation rate (%)	58.9	48.8	51.7	56.7	60.2	60.5	64.9	68.8	72.8	75.2
	Total participants	734	2,519	3,531	4,010	3,703	3,573	3,934	3,838	3,887	3,346
Northern Territory (Total)	Eligible participants	1,322	5,901	7,783	8,151	7,241	6,904	7,072	6,397	5,891	4,811
	Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0	69.5
Australian Capital Territory Divisions	Covariate as Subheading			Summary of data inside data							
	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169	4,394
Canberra(d)	Eligible participants	2,260	6,471	6,448	6,509	5,983	5,805	6,302	5,902	6,044	5,057
	Participation rate (%)	78.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5	86.9
	Total participants	1,477	4,687	5,178	5,786	6,025	5,463	5,191	4,208	3,948	3,485
Fenner(e)	Eligible participants	1,904	6,354	7,121	7,822	7,960	7,155	6,480	5,206	4,692	3,945
	Participation rate (%)	77.6	73.8	72.7	74.0	75.7	76.4	80.1	80.8	84.1	87.8
				NA Yeah							
	Total participants	3,241	9,476	9,995	10,759	10,651	9,916	10,265	9,034	9,117	7,839
Australian Capital Territory (Total)	Eligible participants	4,164	12,825	13,569	14,331	13,943	12,960	12,782	11,108	10,736	9,002
	Participation rate (%)	77.8	73.9	73.7	75.1	76.4	76.5	80.3	81.3	84.9	87.3
Australia											
	Total participants	151,297	438,166	441,658	460,548	462,206	479,360	524,620	517,693	543,449	506,799
Total	Eligible participants	201,439	635,909	646,916	665,250	656,446	660,841	693,850	659,150	664,720	597,386
	Participation rate (%)	75.1	68.9	68.3	69.2	70.4	72.5	75.6	78.5	81.8	84.8

(a) The Federal Electoral Divisions are current as at 24 August 2017
(b) Includes those whose age is unknown — Return of the table junk
(c) Includes Christmas Island and the Cocos (Keeling) Islands
(d) Includes Norfolk Island
(e) Includes Jervis Bay

MS Excel or Die

data

wrangling

tidy data

	area	gender	age	State	Area (sq km)	Eligible participants	Participation rate (%)	Total participants	Total Paticipants
2	Adelaide	Female	18-19 years	SA	76	1341	83.5	1120	1120
3	Adelaide	Female	20-24 years	SA	76	4620	81.2	3750	3750
4	Adelaide	Female	25-29 years	SA	76	4897	81.8	4004	4004
5	Adelaide	Female	30-34 years	SA	76	4784	79.8	3820	3820
6	Adelaide	Female	35-39 years	SA	76	4319	79	3411	3411
7	Adelaide	Female	40-44 years	SA	76	4310	80.6	3472	3472
8	Adelaide	Female	45-49 years	SA	76	4579	81.4	3728	3728
9	Adelaide	Female	50-54 years	SA	76	4475	84.7	3791	3791
10	Adelaide	Female	55-59 years	SA	76	4622	87.3	4033	4033
11	Adelaide	Female	60-64 years	SA	76	4342	89.3	3879	3879
12	Adelaide	Female	65-69 years	SA	76	3970	90.7	3602	3602
13	Adelaide	Female	70-74 years	SA	76	3009	90.3	2716	2716
14	Adelaide	Female	75-79 years	SA	76	2156	88.5	1908	1908
15	Adelaide	Female	80-84 years	SA	76	1673	85.1	1423	1423

https://medium.com/@miles.mcbain/tidying-the-australian-same-sex-marriage-postal-survey-data-with-r-5d35cea07962

Data Vocabulary

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

There are 7 different variables (features) in this spreadsheet. Variables are stored in columns.

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

For each variable, we see there are 4 different observations (samples). Observations are stored in rows.

Demographic Survey Data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Two different types of data

Doctor's Office Measurements Data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	Smith	Jane	65	180	0.60	163
3	4587	Nayef	Mohammed	75	215	1.46	150
4	1727	Doe	Janice	62	124	0.72	177
5	6879	Jordan	Alex	77	160	1.23	205

Tidy Data

1. Each variable you measure should be in a single column

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

2. Every observation of a variable should be in a different row

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

3. There should be one table for each type of data

Demographic Survey Data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Doctor's Office Measurements Data

	A	D	E	F	G
1	ID	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	65	180	0.60	163
3	4587	75	215	1.46	150
4	1727	62	124	0.72	177
5	6879	77	160	1.23	205

4. If you have multiple tables, they should include a column in each *with the same column label* that allows them to be joined or merged

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan					Teacher

Unique Identifier or UID

	A	D	E	F	G
1	ID	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	65	180	0.60	163
3	4587	75	215	1.46	150
4	1727	62	124	0.72	177
5	6879	77	160	1.23	205

Tidy data == rectangular data

A

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints* 5:e3183v1 https://doi.org/10.7287/peerj.preprints.3183v1

Tidy Data Benefits

1. consistent data structure
2. foster tool development
3. require only a small set of tools to be learned
4. allow for datasets to be combined

Good Spreadsheets

Rules for Tidy Spreadsheets

1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Put just one thing in a cell
6. Don't use font color or highlighting as data
7. Save the data as plain text files (i.e. CSV)

1. Be Consistent!

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

Keep exactly the same variable names across spreadsheets.

In these data, sex is always specified as “female” or “male.” Pick a way to code your variables and stick to it.

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints* 5:e3183v1 https://doi.org/10.7287/peerj.preprints.3183v1

2. Choose good names for things

	Do this...	Not This!
Avoid Extra Spaces	'male'	'male '
Use underscores not spaces	doctor_visit_1	Doctor Visit 1
Choose meaningful names	doctor_visit	“F1”

3. Write dates as YYYY-MM-DD

	Do this...	Not This!
Use 'ISO 8601' standard	2018-02-27	*2/27 or* 2_27_2018 *or* Feb 27

4. No empty cells

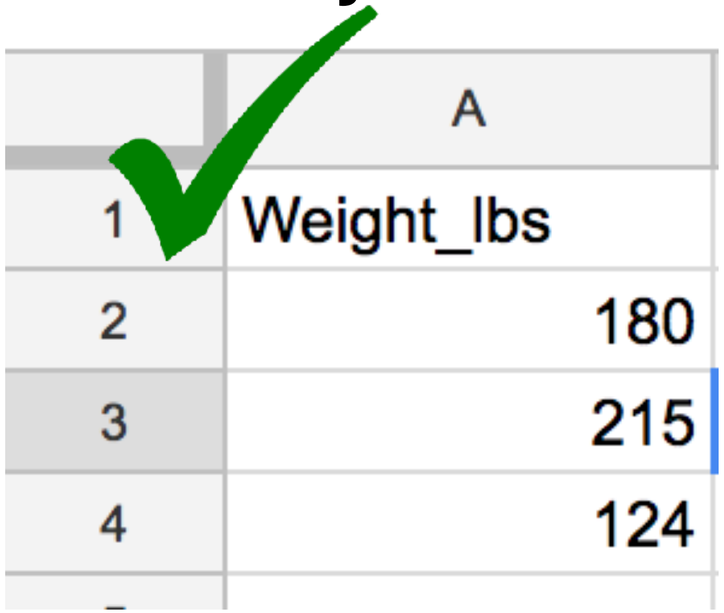
A

	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4

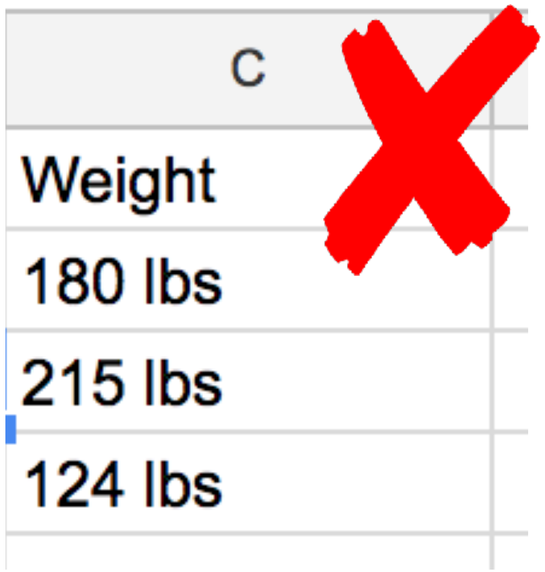
B

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

5. Put just one thing in a cell



	A
1	Weight_lbs
2	180
3	215
4	124



C
Weight
180 lbs
215 lbs
124 lbs

6. Don’t use font color or highlighting as data

A

	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102	2015-06-14	95.3
4	103	2015-06-18	97.5
5	104	2015-06-18	1.1
6	105	2015-06-18	108.0
7	106	2015-06-20	149.0
8	107	2015-06-20	169.4

B

	A	B	C	D
1	id	date	glucose	outlier
2	101	2015-06-14	149.3	FALSE
3	102	2015-06-14	95.3	FALSE
4	103	2015-06-18	97.5	FALSE
5	104	2015-06-18	1.1	TRUE
6	105	2015-06-18	108.0	FALSE
7	106	2015-06-20	149.0	FALSE
8	107	2015-06-20	169.4	FALSE

When..	Be sure to...	So Do this...	Avoid this...	Why?
Naming variables (aka assigning column headers)	Use meaningful variable names	`AgeAtDiagnosis`	`ADx`	`ADx` is an unclear and uninformative abbreviation
Naming variables	Avoid spacing in column headers	`AgeAtDiagnosis`	`Age At Diagnosis`	Spacing in variable names makes the analyst's life more difficult
Naming variables	Use consistent capitalization	`AgeAtDiagnosis`	Using both `AgeAtDiagnosis` and `ageatdiagnosis`	Using consistent column names across tables/spreadsheets simplifies any merging the statistician may have to do.
Naming variables	Avoid using separators, but if it's necessary, use an underscore (`_`)	`IGF1` (or `IGF_1`)	`IGF.1`, `IGF-1`, `IGF/1`, `IGF,1`	Separators (commas, periods, hyphens, slashes, spaces etc.) often have different meanings in coding languages than they do in text. Avoiding them avoids error.
Coding variables	Avoid unnecessary spaces	'male'	'male '	That extra space after 'male ' makes it different from 'male' without a space.
Coding variables	Be consistent!	'male'	'Male','male', and 'M',	In the eyes of the statistician, 'Male','male', and 'M' could be incorrectly perceived as three different values.
Coding variables	Be careful of spelling errors	'male'	'maale'	That extra 'a' makes these two different categories.
Coding date and time	Use ISO 8601 coding	'YYYY-MM-DD'	'MM/DD/YY` and `Month Day, Year`	Consistency simplifies the analyst's life, and YYYY-MM-DD will not be misconstrued if opened in Excel.
Coding missing data	Not leave any cells blank and use a consistent value	'NA'	'0', '-9', red-highlighted blank cells, '.', ' ', '-', ...	Each cell should be filled with a consistent value. Pick a way to denote missingness (ideally 'NA') and stick with it. Avoid using numbers or punctuation to denote missing data.
Entering data	Stick to text and numbers	Convey all information with direct text/numerical entry	Using cell highlighting or font color to convey information	Your analyst may not use the same platform for analysis as you used for data entry, so avoiding font color and cell highlighting will minimize issues.
Generating an Excel file	Save the data in an appropriate format	Use one worksheet per table and save as CSV or text files	Multiple worksheets	Statisticians require this format to import your data onto other platforms.
Entering Data	Avoid entering unnecessary lines of text at the start	Start your first row with variable names	Adding lines of text	This violates the rules of tidy data and makes processing more difficult. Include this information in the "Code book" instead.
Opening files in Excel	Know and avoid its pitfalls	Consistently include one value per cell and be careful of date and time data.	Using macros, splitting cells, and merging cells	These formats are not amenable to data analysis on other platforms.

Common Problems with Messy Data Sets

1. Column headers are values but should be variable names.
2. A single column has multiple variables.
3. Variables have been entered in both rows and columns.
4. Multiple "types" of data are in the same spreadsheet.
5. A single observation is stored across multiple spreadsheets.

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don					10	35
Eva					881	1486
Hin					11	34
Hist					197	223
Jeh					21	30
Jew					30	95

Table 4: ew Forum. Three columns, $75–100k

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Spreadsheet Stud

A

ID	Last	First	Sex	Weight
1004	Smith	Jane	female	170lbs
4587	Nayef	Mohammed	male	200lbs
1727	Doe	Janice	F	120lbs
6879	Jordan	Alex	M	150lbs

B

ID	Last	First	Sex	Weight
1004	Smith	Jane	female	170
4587	Nayef	Mohammed	male	200
1727	Doe	Janice	female	120
6879	Jordan	Alex	male	150

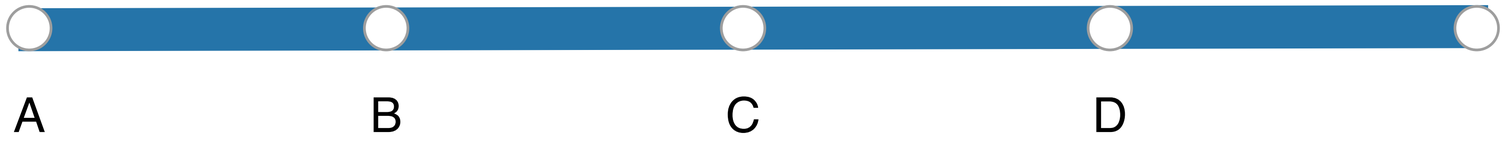
C

ID	Last	First	Sex	Weight
1004	Smith	Jane	female	170
4587	Nayef	Mohammed	male	200
1727	Doe	Janice	F	120
6879	Jordan	Alex	M	150

D

ID	Last	First	Sex	Weight
1004	Smith	Jane	female	170lbs
4587	Nayef	Mohammed	male	200lbs
1727	Doe	Janice	female	120lbs
6879	Jordan	Alex	male	150lbs

Which of these spreadsheets is best?



Spreadsheet Stud

A

ID	Last	First	height_m	height_f
1004	Smith	Jane	NA	65
4587	Nayef	Mohammed	72	NA
1727	Doe	Janice	NA	60
6879	Jordan	Alex	55	NA

B

ID	Last	First	height_m	height_f
1004	Smith	Jane		65
4587	Nayef	Mohammed	72	
1727	Doe	Janice		60
6879	Jordan	Alex	55	

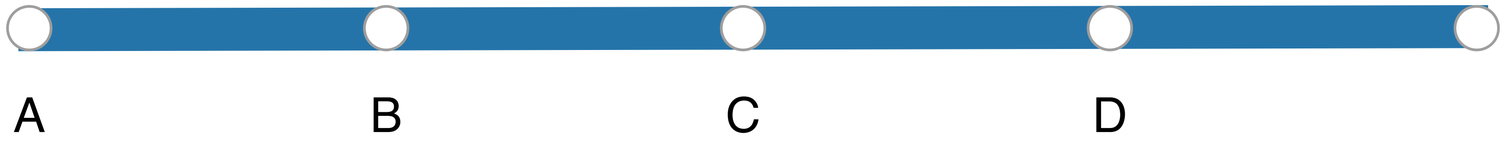
C

ID	Last	First	sex	height
1004	Smith	Jane	female	65
4587	Nayef	Mohammed	male	72
1727	Doe	Janice	fem	60
6879	Jordan	Alex	male	55

D

ID	Last	First	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

Which of these spreadsheets is best?



Clinic Name	Clinic Location	Address	Contact Number	Operational Hours		Services
				Drop-in	By Appointment	
Birth Control and Sexual Health Centre	Dufferin St./Lawrence Ave. W.	Suite 403, 960 Lawrence Ave. W., Toronto, On M6A 3B5	416-789-4541		Monday: 2 pm - 5 pm Tuesday: 4 pm - 7 pm Wednesday: 12 noon - 5 pm Thursday: 5 pm - 8 pm Friday : 12:30 pm - 4:30 pm	Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling Anonymous HIV testing (including the rapid HIV test)
Black Creek Community Health Centre (Sheridan Mall Site)	Jane St./Wilson Ave.	North York Sheridan Mall, Unit 5 2202 Jane St., Toronto, On M2M 1A4	416-249-8000	Tuesday : 4 pm - 6 pm	Tuesday (1pm-4pm	Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
Black Creek Community Health Centre (Yorkgate mall Site)	Jane St./Finch Ave. W.	1 York Gate Blvd., Toronto, On M3N 3A1	416-246-2388	Wednesday: 1 pm - 4:30 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
Special Treatment Clinic	Bay St./College St.	8th Floor, 790 Bay St., Toronto, On M5G 1N8	416-351-3600 ext 2207	Tuesday: 5 pm - 7 pm Wednesday: 5 pm - 7 pm Thursday: 5 pm - 7 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
Taibu Community Health Center	Neilson Rd./McLevin Ave.	27 Tapscott Rd., Toronto, On M1B 4Y7	416-644-3536		Tuesday: 2:30 pm - 6:30 pm	Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling
The Talk Shop	Yonge St./Empress Ave.	5110 Yonge St., Toronto, On M2N 6M1	416-338-7000	Monday: 2 pm - 6:30 pm Wednesday: 9:30 am - 11:30 am, 1 pm - 3:30 pm Thursday: 2 pm - 6:30 pm		Birth control counselling Low cost or free birth control Free condoms Plan B (emergency contraceptive pill) STI testing and free treatment HIV testing Pregnancy testing, counselling and referral Sexuality and relationship counselling Rapid HIV testing

Merged column heading!
Column headings on the second row!

Names/addresses across multiple lines

Attributes spanning multiple rows, without the ID (clinic name) also spanning! Not tidy 🤷‍♀️

NA NA NA NA

NA NA NA NA

NA NA NA NA

NA NA NA NA

https://sharlagelfand.netlify.com/posts/tidying-toronto-open-data/

untidy data

Australian Bureau of Statistics

1800.0 Australian Marriage Law Postal Survey, 2017

Released on 15 November 2017

Table 5 Participation by Federal Electoral Division(a), Males and Age — Gender apartheid

Table junk

Yeah NA

		18-19 years	20-24 years	25-29 years	30-34 years	35-39 years	40-44 years	45-49 years	50-54 years	55-59 years	60-64 years
Lingiari(c)	Total participants	292	1,058	1,465	1,653	1,515	1,516	1,710	1,730	1,753	1,574
	Eligible participants	572	2,910	3,789	3,996	3,607	3,506	3,645	3,331	2,960	2,456
Primary keyndex	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.9	51.9	59.2	64.1
Merged cells				Comma on							
Solomon	Total participants	442	1,461	2,066	2,357	2,188	2,057	2,224	2,108	2,134	1,772
	Eligible participants	750	2,991	3,994	4,155	3,634	3,398	3,427	3,066	2,931	2,355
	Participation rate (%)	58.9	48.8	51.7	56.7	60.2	60.5	64.9	68.8	72.8	75.2
Northern Territory (Total)	Total participants	734	2,519	3,531	4,010	3,703	3,573	3,934	3,838	3,887	3,346
	Eligible participants	1,322	5,901	7,783	8,151	7,241	6,904	7,072	6,397	5,891	4,811
	Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0	69.5
Australian Capital Territory Divisions	Covariate as Subheading			Summary of data inside data							
Canberra(d)	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169	4,394
	Eligible participants	2,260	6,471	6,448	6,509	5,983	5,805	6,302	5,902	6,044	5,057
	Participation rate (%)	78.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5	86.9
Fenner(e)	Total participants	1,477	4,687	5,178	5,786	6,025	5,463	5,191	4,208	3,948	3,465
	Eligible participants	1,904	6,354	7,121	7,822	7,960	7,155	6,480	5,206	4,692	3,945
	Participation rate (%)	77.6	73.8	72.7	74.0	75.7	76.4	80.1	80.8	84.1	87.8
				NA Yeah							
Australian Capital Territory (Total)	Total participants	3,241	9,476	9,995	10,759	10,651	9,916	10,265	9,034	9,117	7,859
	Eligible participants	4,164	12,825	13,569	14,331	13,943	12,960	12,782	11,108	10,736	9,002
	Participation rate (%)	77.8	73.9	73.7	75.1	76.4	76.5	80.3	81.3	84.9	87.3
Australia											
Total	Total participants	151,297	438,166	441,658	460,548	462,206	479,360	524,620	517,693	543,449	506,799
	Eligible participants	201,439	635,909	646,916	665,250	656,446	660,841	693,850	659,150	664,720	597,386
	Participation rate (%)	75.1	68.9	68.3	69.2	70.4	72.5	75.6	78.5	81.8	84.8

(a) The Federal Electoral Divisions are current as at 24 August 2017
(b) Includes those whose age is unknown
(c) Includes Christmas Island and the Cocos (Keeling) Islands
(d) Includes Norfolk Island
(e) Includes Jervis Bay

Return of the table junk

MS Excel or Die

data wrangling

tidy data

	area	gender	age	State	Area (sq km)	Eligible participants	Participation rate (%)	Total participants	Total Paticipants
2	Adelaide	Female	18-19 years	SA	76	1341	83.5	1120	1120
3	Adelaide	Female	20-24 years	SA	76	4620	81.2	3750	3750
4	Adelaide	Female	25-29 years	SA	76	4897	81.8	4004	4004
5	Adelaide	Female	30-34 years	SA	76	4784	79.8	3820	3820
6	Adelaide	Female	35-39 years	SA	76	4319	79	3411	3411
7	Adelaide	Female	40-44 years	SA	76	4310	80.6	3472	3472
8	Adelaide	Female	45-49 years	SA	76	4579	81.4	3728	3728
9	Adelaide	Female	50-54 years	SA	76	4475	84.7	3791	3791
10	Adelaide	Female	55-59 years	SA	76	4622	87.3	4033	4033
11	Adelaide	Female	60-64 years	SA	76	4342	88.3	3879	3879
12	Adelaide	Female	65-69 years	SA	76	3970	90.7	3602	3602
13	Adelaide	Female	70-74 years	SA	76	3009	90.3	2716	2716
14	Adelaide	Female	75-79 years	SA	76	2156	88.5	1908	1908
15	Adelaide	Female	80-84 years	SA	76	1673	85.1	1423	1423

https://medium.com/@miles.mcbain/tidying-the-australian-same-sex-marriage-postal-survey-data-with-r-5d35cea07962

Tidy Data Benefits

1. consistent data structure
2. foster tool development
3. require only a small set of tools to be learned
4. allow for datasets to be combined

TIDY data is NOT the same as CLEAN data

Common data wrangling tasks

(and verbs)

- subset dataset
  - filter / query: filter rows
  - select: select columns
  - slice: take a continuous subset
- change order
  - arrange / sort: change order of rows
  - reorder / sort: change order of columns
- add columns
  - join: merge columns from two tables
  - mutate: new column based on existing column(s)
- summarize data
  - groupby: group by other variables
  - summarize / aggregate / apply: reduce multiple values down to a single value

GBBO: Great British Bake Off

- Baking competition for amateur bakers
- Filmed in gardens
- Weekly eliminations
- Each week:
  - Signature Challenge - show off their tried-and-tested recipes
  - Technical Challenge - produce a certain finished product when given only limited – or even minimal – instructions.
  - Showstopper Challenge - show off their skills and talent; professional appearance & outstanding flavours.
- 10 seasons since 2010

What types of information (data) would this show generate?

https://en.wikipedia.org/wiki/The_Great_British_Bake_Off

ratings results challenges season bakers series

series	episode	baker	signature
1	2	Jasminder	Millionaires
1	2	Jonathan	Honey and C
1	2	Louise	Stained Gla
1	2	Miranda	Fresh Vanill
2			
2			

series	episodes	premiere	finale	winner	avg_uk_viewers	day_of_week	timeslot	channel	runner_up_1	runner_up_2
1	6	8/17/10	9/21/10	Edd Kimber	2.77	Tuesday	8:00pm	BBC Two	Miranda Gore Browne	Ruth Clemens
2	8	8/14/11	10/4/11	Joanne Whea	4	Tuesday	8:00pm	BBC Two	Holly Bell	Mary-Anne Boermans
3	10	8/14/12	10/16/12	John Whaite	5	Tuesday	8:00pm	BBC Two	Brendan Lynch	James Morton
4	10	8/20/13	10/22/13	Frances Quir	7.35	Tuesday	8:00pm	BBC Two	Kimberley Wilson	Ruby Tandoh
5	10	8/6/14	10/8/14	Nancy Birtwl	10.04	Wednesday	8:00pm	BBC One	Luis Troyano	Richard Burr
6	10	8/5/15	10/7/15	Nadiya Hussa	12.5	Wednesday	8:00pm	BBC One	Ian Cumming	Tamal Ray
7	10	8/24/16	10/26/16	Candice Brov	13.85	Wednesday	8:00pm	BBC One	Andrew Smyth	Jane Beedle
8	10	8/29/17	10/31/17	Sophie Faldo	9.29	Tuesday	8:00pm	Channel 4	Kate Lyon	Steven Carter-Bailey

series	baker_full	age	occupation	hometown
1	Ruth Clemens	31	Retail manager/Housewife	Poynton, Cheshire
2	Ben Frazer	31	Graphic Designer	Northampton
2	Holly Bell	31	Advertising executive	Leicester
2	Ian Vallance	40	Fundraiser for English Heritage	Dunstable, Bedfordshire
2	Janet Basu	63	Teacher of Modern Languages	Formby, Liverpool

1	1		
1	2		
1	3		
1	4		
1	5		
1	6		

5	3	2	20-Sep-15
5	4	2	27-Sep-15
5	5	2	4-Oct-15
5	6	2	11-Oct-15
5	7	2	18-Oct-15
5	8	2	25-Oct-15
5	9	2	1-Nov-15
5	10	2	8-Nov-15

How would you get information about all GBBO challenges?

Contains information about what was baked for each challenge in each episode

challenges
series
episode
baker
signature
technical
showstopper

`challenges` and `results` are linked by `series`, `episode`, & `baker`

results
series
episode
baker
result

Contains information about who won in each episode

challenges

series	episode	baker	signature	technical	showstopper
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche

results

series	episode	baker	result
1	1	Annetha	IN
1	2	Annetha	OUT
1	6	Edd	WINNER
2	8	Ben	NA
2	8	Holly	RUNNER-UP

Inner Join: include any rows in *both*

challenges

series	episode	baker	signature	technical	showstopper
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche

results

series	episode	baker	result
1	1	Annetha	IN
1	2	Annetha	OUT
1	6	Edd	WINNER
2	8	Ben	NA
2	8	Holly	RUNNER-UP

Inner Join: include any rows in *both*

tables

challenges

series	episode	baker	signature	technical	showstopper
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche

results

series	episode	baker	result
1	1	Annetha	IN
1	2	Annetha	OUT
1	6	Edd	WINNER
2	8	Ben	NA
2	8	Holly	RUNNER-UP

series	episode	baker	signature	technical	showstopper	result
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream	IN
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches	WINNER

Left Join: include all rows in *first* table

challenges

series	episode	baker	signature	technical	showstopper
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche

results

series	episode	baker	result
1	1	Annetha	IN
1	2	Annetha	OUT
1	6	Edd	WINNER
2	8	Ben	NA
2	8	Holly	RUNNER-UP

Left Join: include all rows in *first* table

`left join`

series	episode	baker	signature	technical	showstopper	result
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream	IN
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches	WINNER
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels	NA
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche	NA

FILL IN MISSING INFORMATION WITH NAs

Right Join: include all rows in *second*

challenges

series	episode	baker	signature	technical	showstopper
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche

results

series	episode	baker	result
1	1	Annetha	IN
1	2	Annetha	OUT
1	6	Edd	WINNER
2	8	Ben	NA
2	8	Holly	RUNNER-UP

`right join`

Right Join: include all rows in *second* table

series	episode	baker	signature	technical	showstopper	result
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream	IN
1	2	Annetha	NA	NA	NA	OUT
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches	WINNER
2	8	Ben	NA	NA	NA	NA
2	8	Holly	NA	NA	NA	RUNNER-UP

FILL IN MISSING INFORMATION WITH NAs

Full Join: include any row in *either*

challenges

series	episode	baker	signature	technical	showstopper
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche

results

series	episode	baker	result
1	1	Annetha	IN
1	2	Annetha	OUT
1	6	Edd	WINNER
2	8	Ben	NA
2	8	Holly	RUNNER-UP

Full Join: include any row in *either* table

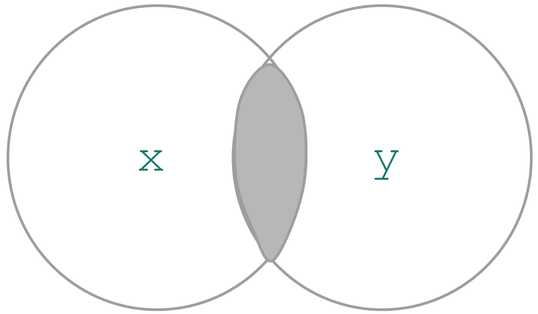
`full join`

series	episode	baker	signature	technical	showstopper	result
1	1	Annetha	Light Jamaican Black Cake w/ Strawberries and Cream	2	Red, White & Blue Chocolate Cake with Cigarellos, Fresh Fruit, and Cream	IN
1	2	Annetha	NA	NA	NA	OUT
1	6	Edd	Cinnamon and Banana Cake	NA	24 Chocolate and Ginger Tarts 24 Lemon Scones w/ Passion Fruit Curd 24 Raspberry Choux Buns 24 Finger Sandwiches	WINNER
1	3	Ruth	Maple and Pecan Bread	2	Chocolate and Orange Panettone Cranberry Bagels	NA
2	6	Holly	Father Christmas' Baked Cheesecake	3	Hansel and Gretel's Croquembouche	NA
2	8	Ben	NA	NA	NA	NA
2	8	Holly	NA	NA	NA	RUNNER-UP

FILL IN MISSING INFORMATION WITH NAs

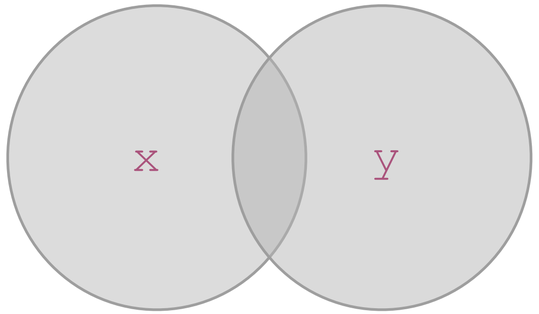
inner

Include any row in both tables



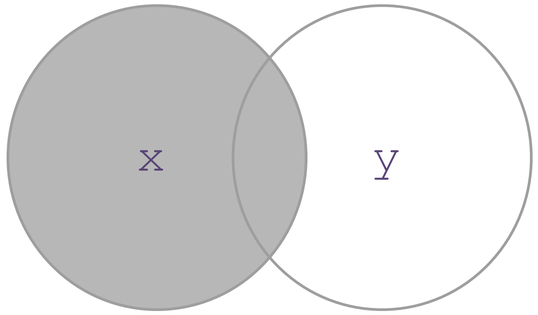
full

Include any row in either table



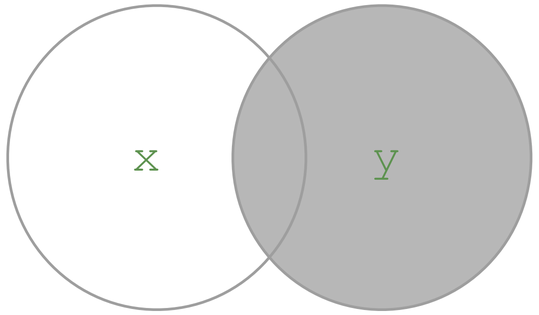
left

Include all rows in 1st table



right

Include all rows in 2nd table



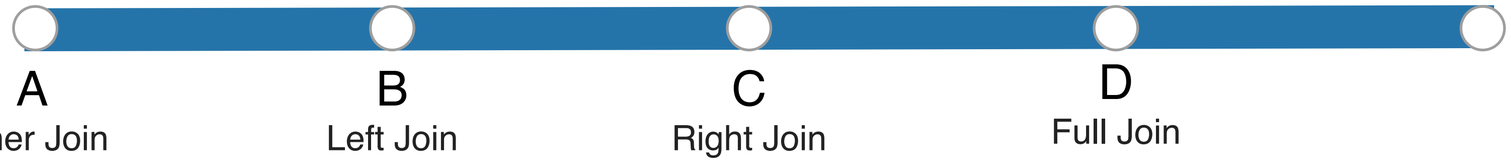
Join Time

What's going on here (we're matching on ID) ?

character	ID	gender	adult	appearance
Eleven	E17W	F	0	main
Will Byers	W14B	M	0	main
MIke Wheeler	M24W	M	0	main
Lucas Sinclair	L46S	NB	0	main
Dustin Henderson	D99H	M	0	main
Nancy Wheeler	N71W	F	0	main
Barbara Holland	B73H	F	1	recurring
Jim Hopper	J19H	M	1	main
Scott Clarke	S28C	M	1	recurring

characteristic	ID
psychic	E17W
telekinetic	E17W
vanishing	W14B
leader	M24W
skeptic	L46S
leader	S28C
skeptic	M22X

character	ID	gender	adult	appearance	characteristic
Eleven	E17W	F	0	main	psychic
Eleven	E17W	F	0	main	telekinetic
Will Byers	W14B	M	0	main	vanishing
MIke Wheeler	M24W	M	0	main	leader
Lucas Sinclair	L46S	NB	0	main	skeptic
Scott Clarke	S28C	M	1	recurring	leader
NA	M22X	NA	NA	NA	skeptic



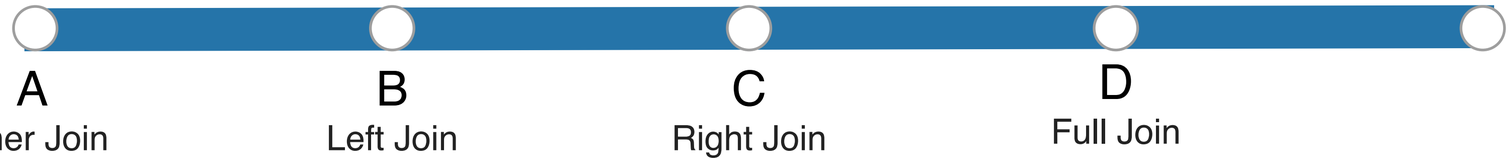
Join Time

What's going on here (matching on ID)?

character	ID	gender	adult	appearance
Eleven	E17W	F	0	main
Will Byers	W14B	M	0	main
MIke Wheeler	M24W	M	0	main
Lucas Sinclair	L46S	NB	0	main
Dustin Henderson	D99H	M	0	main
Nancy Wheeler	N71W	F	0	main
Barbara Holland	B73H	F	1	recurring
Jim Hopper	J19H	M	1	main
Scott Clarke	S28C	M	1	recurring

characteristic	ID
psychic	E17W
telekinetic	E17W
vanishing	W14B
leader	M24W
skeptic	L46S
leader	S28C
skeptic	M22X

character	ID	gender	adult	appearance	characteristic
Eleven	E17W	F	0	main	psychic
Eleven	E17W	F	0	main	telekinetic
Will Byers	W14B	M	0	main	vanishing
MIke Wheeler	M24W	M	0	main	leader
Lucas Sinclair	L46S	NB	0	main	skeptic
Dustin Henderson	D99H	M	0	main	NA
Nancy Wheeler	N71W	F	0	main	NA
Barbara Holland	B73H	F	1	recurring	NA
Jim Hopper	J19H	M	1	main	NA
Scott Clarke	S28C	M	1	recurring	leader
NA	M22X	NA	NA	NA	skeptic



Spelling out why having tidy data helps use these data manipulations

filter rows

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

id	last	first	sex	height
1004	Smith	Jane	F	65
1727	Doe	Janice	F	60

```
filter(sex == 'F')
```

filter rows

An example of why tidy data are what you want to work with.

id	last	first	height_m	height_f
1004	Smith	Jane	None	65
4587	Nayef	Mohammed	72	None
1727	Doe	Janice	None	60
6879	Jordan	Alex	55	None

The wrong answer from the clicker question earlier

```
filter(sex == 'F')
```

filter rows

An example of why tidy data are what you want to work with.

id	last	first	height_m	height_f
1004	Smith	Jane	None	65
4587	Nayef	Mohammed	72	None
1727	Doe	Janice	None	60
6879	Jordan	Alex	55	None

The wrong answer from the clicker question earlier

```
filter(sex == 'F')
```

arrange rows

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

```
arrange(height)
```

arrange rows

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

id	last	first	sex	height
6879	Jordan	Alex	M	55
1727	Doe	Janice	F	60
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72

```
arrange(height)
```

arrange rows

An example of why tidy data are what you want to work with.

id	last	first	height_m	height_f
1004	Smith	Jane	None	65
4587	Nayef	Mohammed	72	None
1727	Doe	Janice	None	60
6879	Jordan	Alex	55	None

The wrong answer from the clicker question earlier

```
arrange(height_m, height_f)
```

arrange rows

An example of why tidy data are what you want to work with.

id	last	first	height_m	height_f
1004	Smith	Jane	None	65
4587	Nayef	Mohammed	72	None
1727	Doe	Janice	None	60
6879	Jordan	Alex	55	None

id	last	first	height_m	height_f
6879	Jordan	Alex	55	None
4587	Nayef	Mohammed	72	None
1727	Doe	Janice	None	60
1004	Smith	Jane	None	65

The wrong answer from the clicker question earlier

```
arrange(height_m, height_f)
```

select & reorder columns

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

```
select(id, first, sex)
```

select & reorder columns

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

id	first	sex
1004	Jane	F
4587	Mohammed	M
1727	Janice	F
6879	Alex	M

```
select(id, first, sex)
```

select & reorder columns

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

first	sex	id
Jane	F	1004
Mohammed	M	4587
Janice	F	1727
Alex	M	6879

```
select(first, sex, id)
```

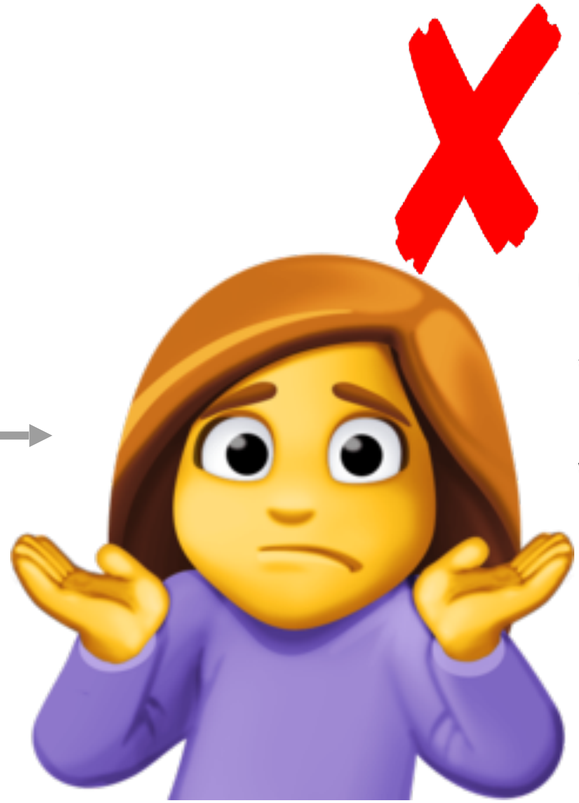
You can change up the order of the variables to reorder columns

select & reorder columns

id	last	first	height_m	height_f
1004	Smith	Jane	None	65
4587	Nayef	Mohammed	72	None
1727	Doe	Janice	None	60
6879	Jordan	Alex	55	None

```
select(first, sex, id)
```

The wrong answer from the clicker question earlier



An example of why tidy data are what you want to work with.

mutate to create a new column

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

```
mutate(patient_dr = 'Grey' if id < 2000 else 'Shepherd')
```

mutate to create a new column

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

id	last	first	sex	height	patient_dr
1004	Smith	Jane	F	65	Grey
4587	Nayef	Mohammed	M	72	Shepherd
1727	Doe	Janice	F	60	Grey
6879	Jordan	Alex	M	55	Shepherd

```
mutate(patient_dr = 'Grey' if id < 2000 else 'Shepherd')
```

Grouping & Summarizing

group by and summarize

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

```
summarize('height', mean)
```

group by and summarize

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

height
63

```
summarize('height', mean)
```

group by and summarize

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

```
group by('sex')
summarize(mean)
```

group by and summarize

id	last	first	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

sex	height
F	62.5
M	63.5

```
group by(sex)
summarize('height', mean)
```

group by and summarize

id	last	first	height_m	height_f
1004	Smith	Jane	NA	65
4587	Nayef	Mohammed	72	NA
1727	Doe	Janice	NA	60
6879	Jordan	Alex	55	NA

The wrong answer from the clicker question earlier

```
group by(sex)
summarize('height', mean)
```

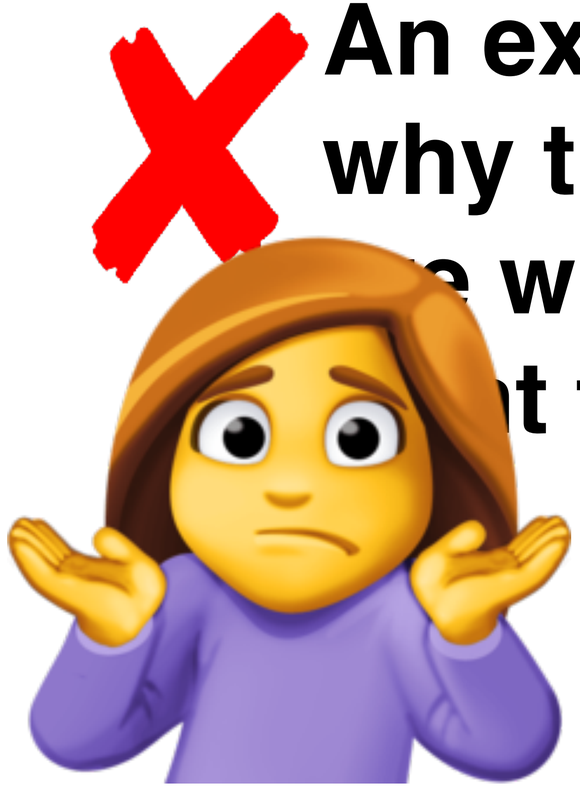
An example of why tidy data are what you want to work with.

group by and summarize

id	last	first	height_m	height_f
1004	Smith	Jane	NA	65
4587	Nayef	Mohammed	72	NA
1727	Doe	Janice	NA	60
6879	Jordan	Alex	55	NA

The wrong answer from the clicker question earlier

```
group by(sex)
summarize('height', mean)
```

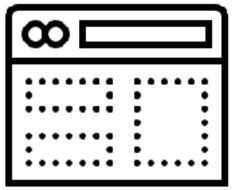


An example of why tidy data e what you t to work

Unstructured Data

Unstructured Data Types

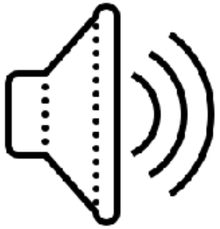
Text files and documents



Websites and applications

Sensor data

Image files



Audio files

Video files

Email data

Social media data

Data Structures Review

Structured data

- can be stored in database SQL
- tables with rows and columns
- requires a relational key
- 5-10% of all data

Semi-structured data

- doesn't reside in a relational database
- has organizational properties (easier to analyze)
- CSV, XML, JSON

Unstructured

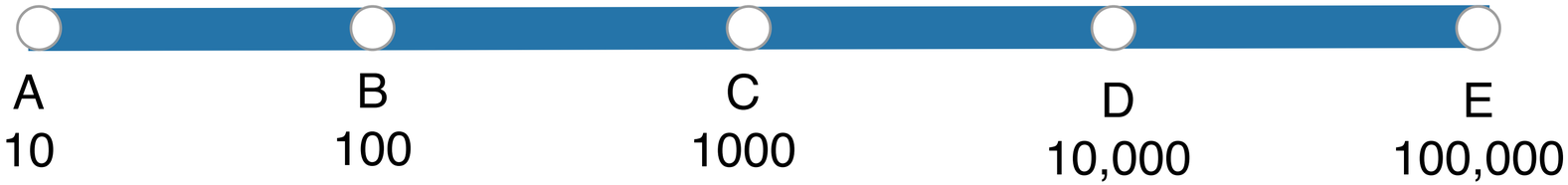
- non-tabular data
- 80% of the world's data
- images, text, audio, videos

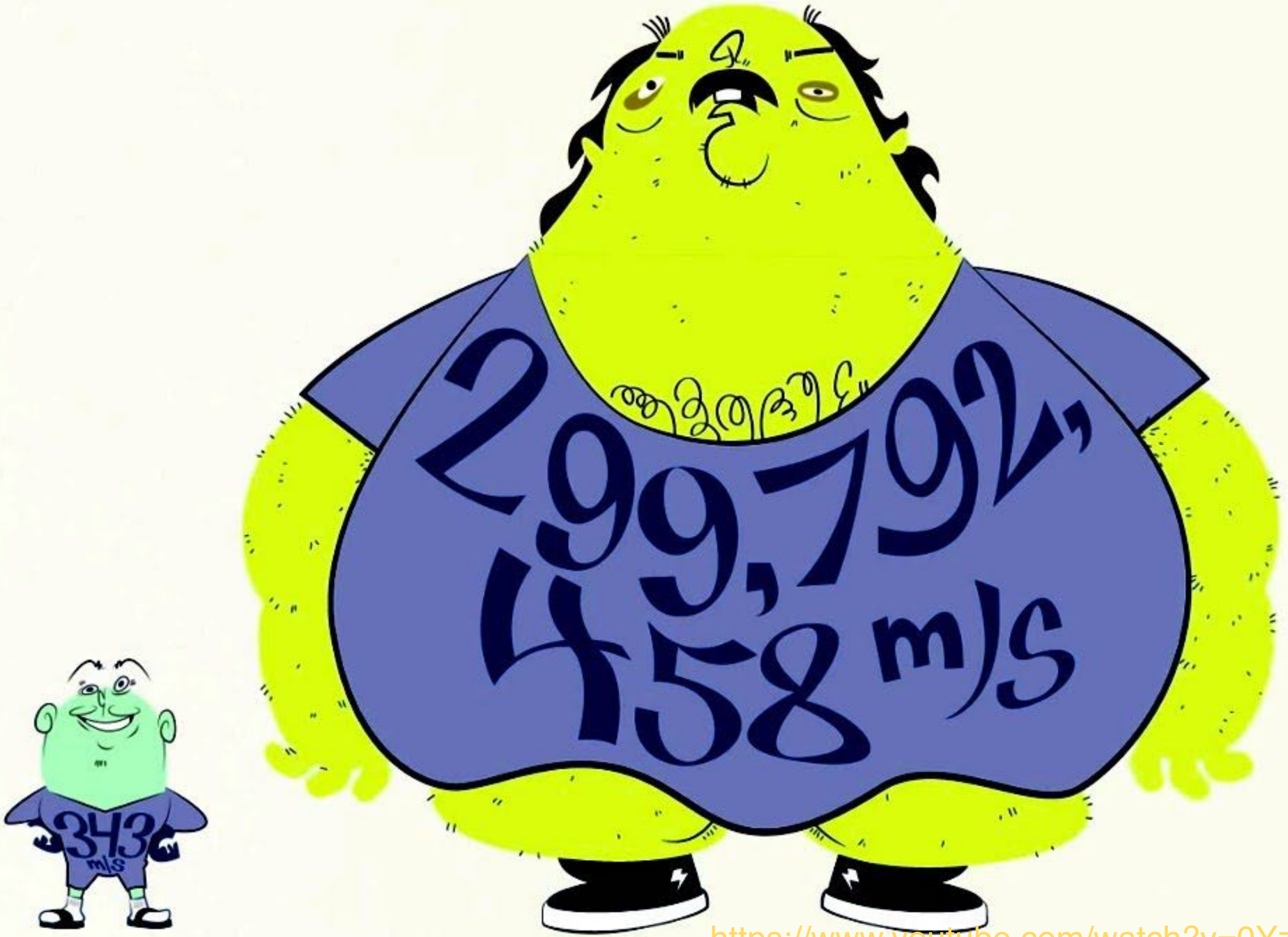
Data Intuition

Fermi Estimation

https://forms.gle/C982naWtU9RvHqAb7

Approximately how many piano tuners do you think there are in the city of Chicago?





https://www.youtube.com/watch?v=0YzvupOX8Is

Has humanity produced enough paint to cover the entire land area of the Earth?

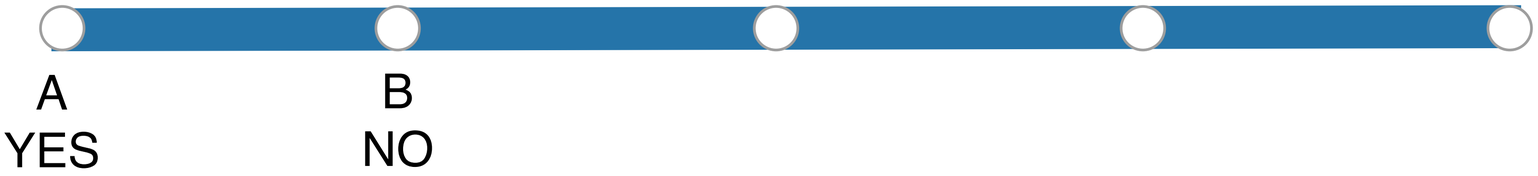
—Josh (Bolton, MA)

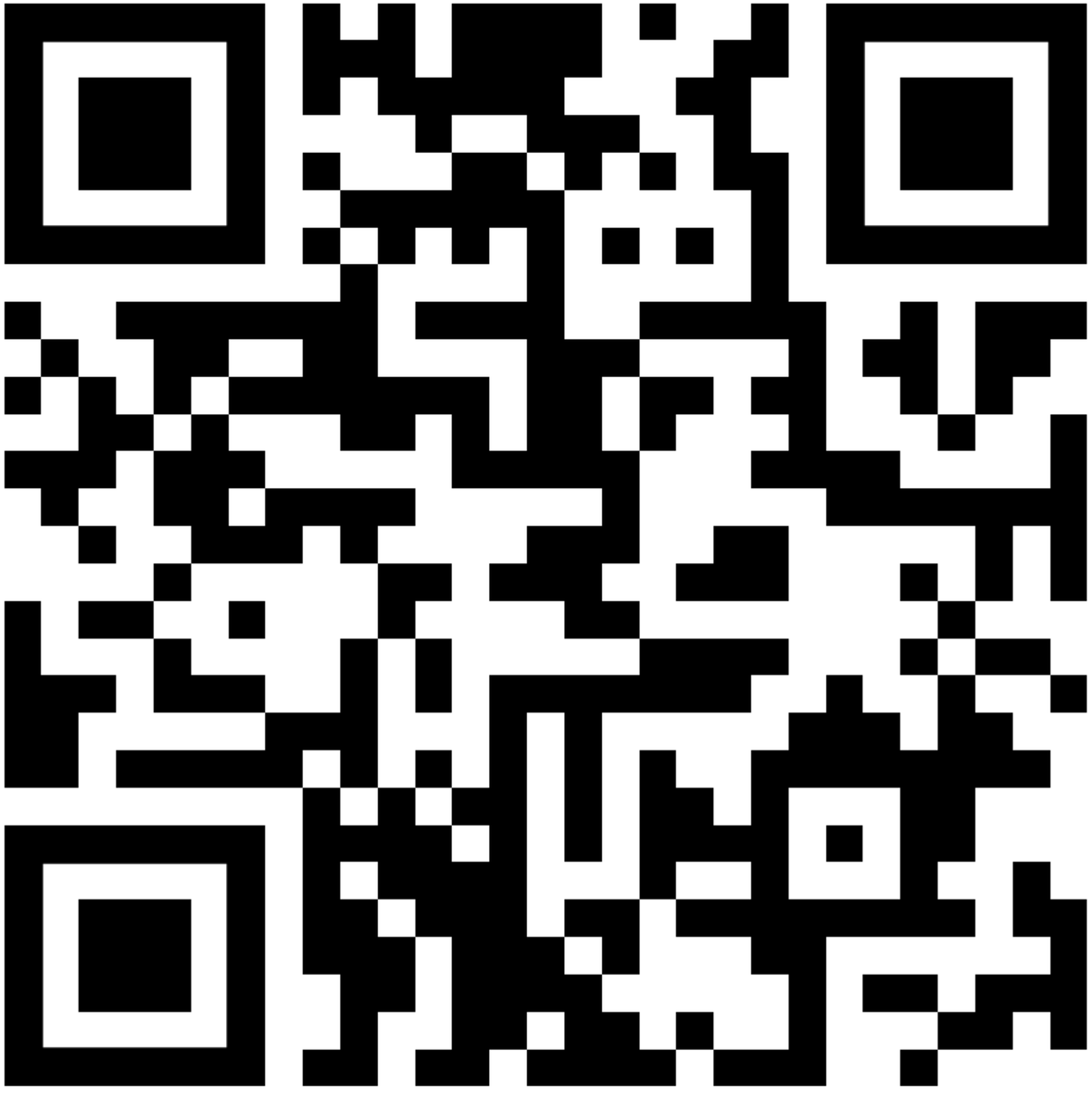
https://what-if.xkcd.com/84/

Fermi Estimation

https://forms.gle/shS84W1tai4SDrVF9

Has humanity produced enough paint to cover the entire land area of the Earth?

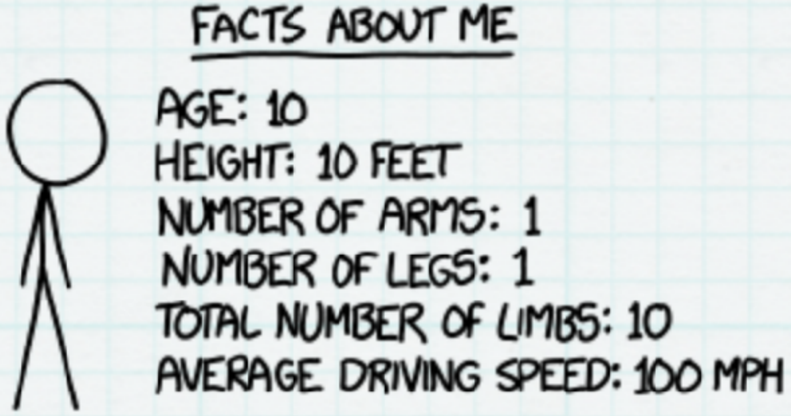




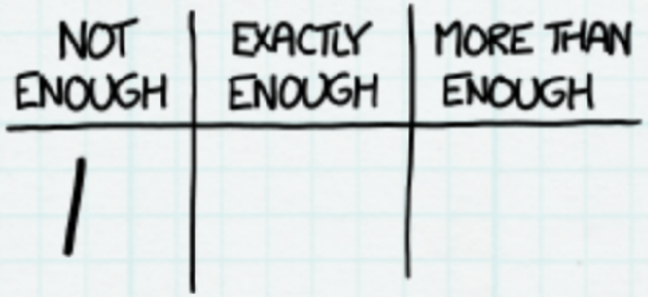
This answer is pretty straightforward. We can look up the size of the world's paint industry, extrapolate backward to figure out the total amount of paint produced. We'd also need to make some assumptions about how we're painting the ground. Note: When we get to the Sahara desert, I recommend not using a brush.

https://what-if.xkcd.com/84/

But first, let's think about different ways we might come up with a guess for what the answer will be. In this kind of thinking—often called **Fermi estimation**—all that matters is getting in the right ballpark; that is, the answer should have about the right number of digits. In Fermi estimation, you can round [1] all your answers to the nearest order of magnitude:



Let's suppose that, on average, everyone in the world is responsible for the existence of two rooms, and they're both painted. My living room has about 50 square meters of paintable area, and two of those would be 100 square meters. 7.15 billion people times 100 square meters per person is a little under a trillion square meters —an area smaller than Egypt.



https://what-if.xkcd.com/84/

Let's make a wild guess that, on average, one person out of every thousand spends their working life painting things. If I assume it would take me three hours to paint the room I'm in,[2] and 100 billion people have ever lived, and each of them spent 30 years painting things for 8 hours a day, we come up with 150 trillion square meters ... just about exactly the land area of the Earth.

How much paint does it take to paint a house? I'm not enough of an adult to have any idea, so let's take another Fermi guess.

Based on my impressions from walking down the aisles, home improvement stores stock about as many light bulbs as cans of paint. A normal house might have about 20 light bulbs, so let's assume a house needs about 20 gallons of paint.[3] Sure, that sounds about right.

The average US home costs about $200,000. Assuming each gallon of paint covers about 300 square feet, that's a square meter of paint per $300 of real estate. I vaguely remember that the world's real estate has a combined value of something like $100 trillion,[4] which suggests there's about 300 billion square meters of paint on the world's real estate. That's about one New Mexico.

Of course, both of the building-related guesses could be overestimates (lots of buildings are not painted) or underestimates (lots of things that are not buildings [5] are painted) But from these wild Fermi estimates, my guess would be that there probably isn't enough paint to cover all the land.

So, how did Fermi do?

According to the report **The State of the Global Coatings Industry**, the world produced 34 billion liters of paints and coatings in 2012.

There's a neat trick that can help us here. If some quantity—say, the world economy—has been growing for a while at an annual rate of **n**—say, 3% (0.03)—then the most recent year's share of the whole total so far is $1 - \frac{1}{1+n}$, and the whole total so far is the most recent year's amount times $1 + \frac{1}{n}$.

If we assume paint production has, in recent decades, followed the economy and grown at about 3% per year, that means the total amount of paint produced equals the current yearly production times 34.[6] That comes out to a little over a trillion liters of paint. At 30 square meters per gallon,[7] that's enough to cover 9 trillion square meters—about the area of the United States.

So the answer is no; there's not enough paint to cover the Earth's land, and—at this rate—probably won't be enough until the year 2100.

Data Intuition

1. Think about your question and your expectations
2. Do some Fermi calculations (back of the envelope calculations)
3. Write code & look at outputs <- think about those outputs
4. Use your gut instinct / background knowledge to guide you
5. Review code & fix bugs

On your own (meaning w/o Googling), please fill out quickly:

https://forms.gle/CREcpMkYDLYTUp2s6