Do you want to learn more about sports analytics? Do you want to hear from sports industry professionals?

If so, join Triton Ball! UCSD’s sports analytics club.

TRITON BALL

SPORTS ANALYTICS CLUB

https://www.instagram.com/tritonball_ucsd?igsh=MTFkM2h5NzBsbmxnNw%3D%3D&utm_source=qr

TRITONBALL_UCSD

https://discord.gg/JcTP2xXBej
Discord link is also in instagram Bio

Logistics

- Due Monday
  - Quiz 1
- Due Wed
  - Pre course survey for EC
    - Section 1 ==> EC
    - Section 2 ==> Participation in the pedagogical experiment
  - Practice Assignment on Datahub (1% of grade)
- Due Fri
  - D1 on Datahub
  - #FinAid quiz on Canvas
- Discussion Lab this week:
  - Help with D1
  - Help with git usage and software installation
    - You may find it very helpful to install (if you don’t already have!)
      - https://git-scm.com/downloads
      - https://www.anaconda.com/download

Data
tidiness & intuition

Jason G. Fleischer, Ph.D
UC San Diego

Department of Cognitive Science
jfleischer@ucsd.edu
https://jgfleischer.com

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

Structured Data

Databases!

What is a DB?? An organized collection of structured information

Manage huge datasets

Control access to data

Allow users to find a subset of data with a query

Run an analysis on data inside the DB and return a report

Structured Data

Examples of relational DB

- *SQLite*
- *MySQL*
- *Postgres*

Relational DB work using tables of data with "relationships" established between tables

Examples of non-relational DB

- *Hadoop*
- *Hive*
- *Apache CouchBase*

NoSQL DB work with key-value pairs to lookup data, and that's exactly like JSON slides coming up.

(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

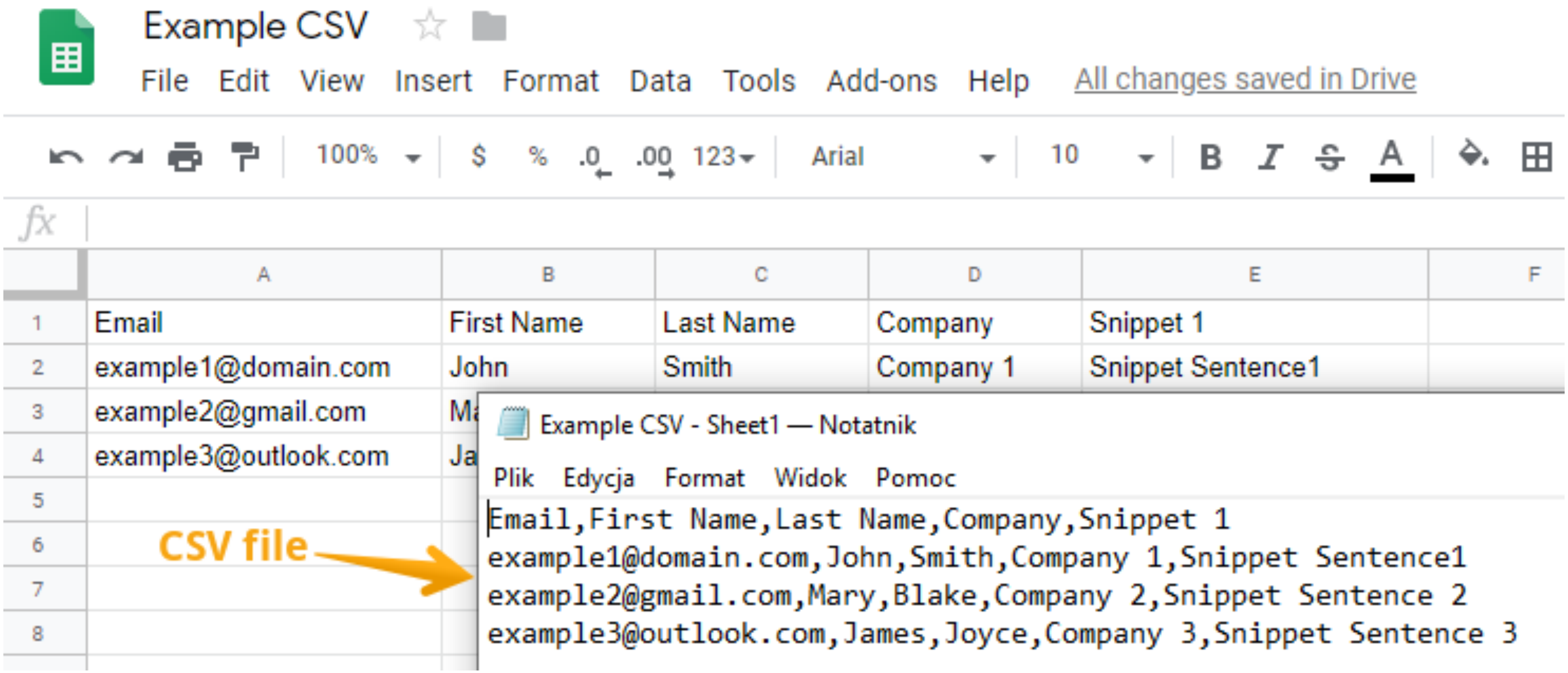
Has the extension ".csv"

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
```

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 B I S A
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
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

{"Name": "Isabela"}

key

value

JSON

These are all nested within `attributes`

These are all nested within "`Good For`"

```
"attributes": {
  "Take-out": true,
  "Wi-Fi": "free",
  "Drive-Thru": true,
  "Good For": {
    "dessert": false,
    "latenight": false,
    "lunch": false,
    "dinner": false,
    "breakfast": false,
    "brunch": false
  },
```

JSON

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


```json
{
 "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

is

{json}

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')"
 ]
},
```

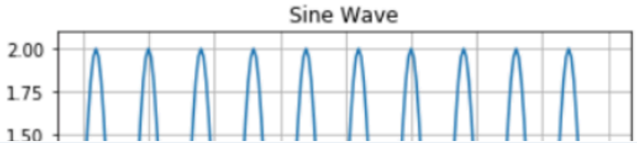
In [10]:
```python
import numpy as np
import matplotlib.pyplot as plt

# 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+80fVdXdmOnl1q17ezBm/T6f+QDdXVXnVtU996z3HFFKESNGjBgxVgZigREjRow
YMbQQC4wYMWLEiBGWFWGDEiBEjRgwtxAIjRowYMWJoIRYYMWLEiBFDC7HAiBEjRowYMWLEiBFDC7HAiBEDEJG/JiKPL/c4YsQ4nxELjBgfGojIXSLyoojMiMg
```

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:

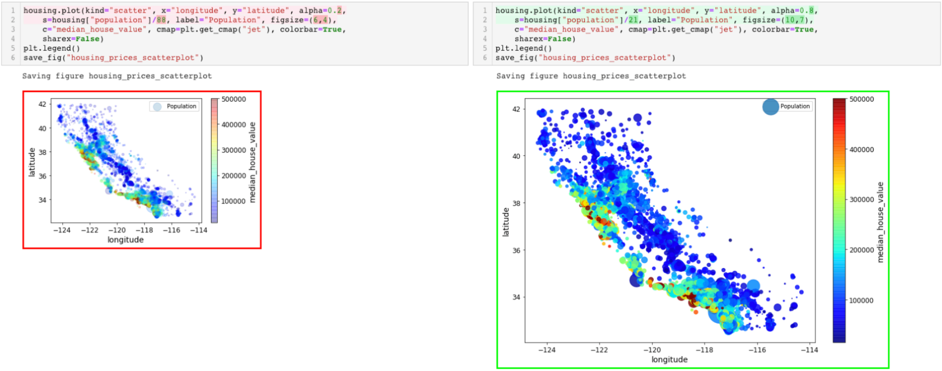
- It is a manual process.
- Collaborators on other machines will need to rerun the notebook to see the output, requiring additional time and setup.

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

A **node**

```
$node
<tag>
    <tag2> more content </
tag2>
    <tag3> more content </
tag3>
</tag>
```

An *opening* **tag**

An **element**

A *closing* **tag**

XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<customers>
    <customer>
        <customer_id>1</customer_id>
        <first_name>John</first_name>
        <last_name>Doe</last_name>
        <email>john.doe@example.com</email>
    </customer>
    <customer>
        <customer_id>2</customer_id>
        <first_name>Sam</first_name>
        <last_name>Smith</last_name>
        <email>sam.smith@example.com</email>
    </customer>
    <customer>
        <customer_id>3</customer_id>
        <first_name>Jane</first_name>
        <last_name>Doe</last_name>
        <email>jane.doe@example.com</email>
    </customer>
</customers>
```


XML

Relational Databases: A set of interdependent tables

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

Results | Messages

	EmployeeId	FirstName	LastName	DepartmentName
1	1	Ken	Sanchez	Executive
2	2	Terri	Duffy	Engineering
3	3	Roberto	Tamburello	Engineering
4	4	Rob	Walters	Engineering
5	5	Gail	Erickson	Engineering
6	6	Jossef	Goldberg	Engineering
7	7	Dylan	Miller	Support
8	8	Diane	Margheim	Support
9	9	Gigi	Matthew	Support
10	10	Michael	Raheem	Support

relational database

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

relational database

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

id	inspection_ date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

relational database

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

id	inspection_ date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

Two different restaurants with the same name will have different unique identifiers

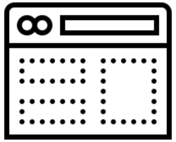
relational database

Unstructured Data

Some datasets record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.

Unstructured Data Types

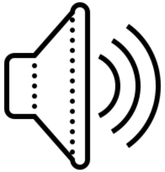
Text files and documents



Websites and applications

Sensor data

Image files

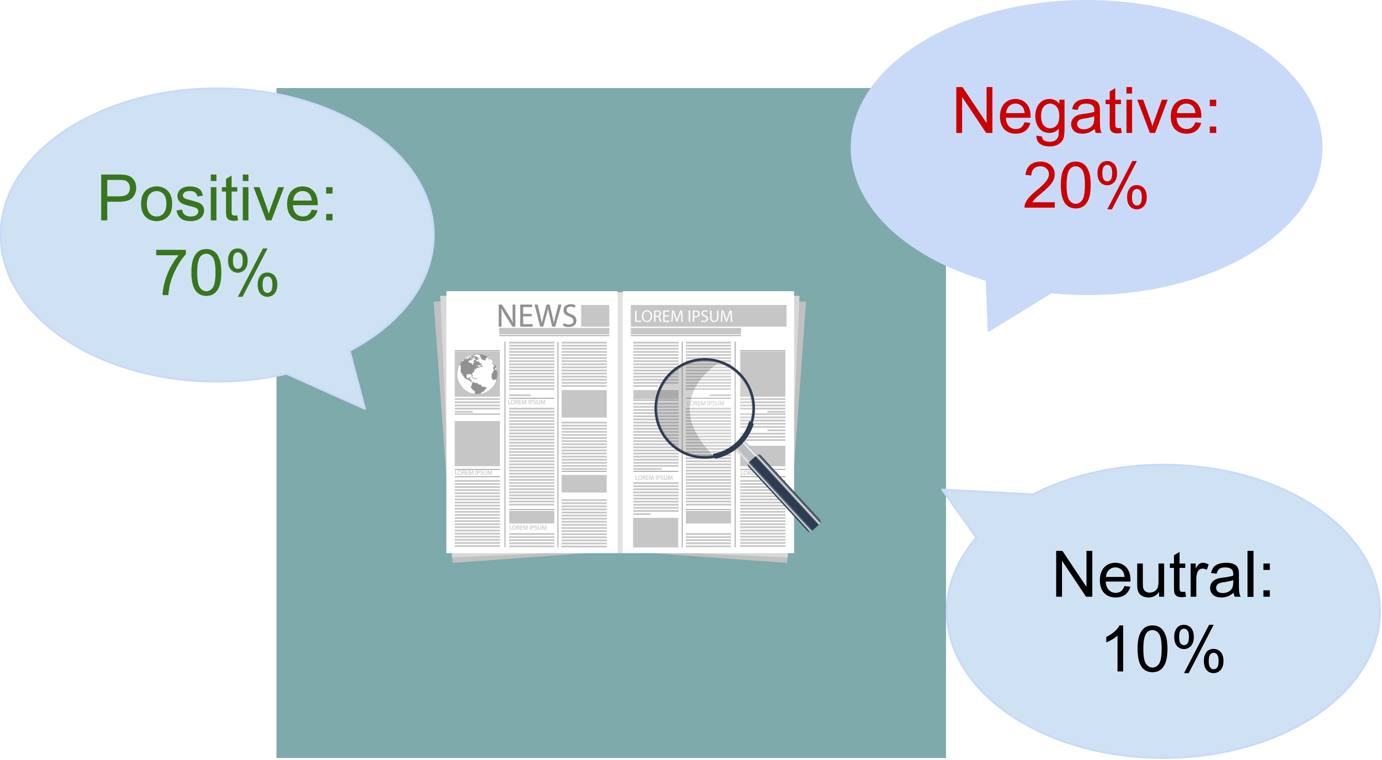


Audio files

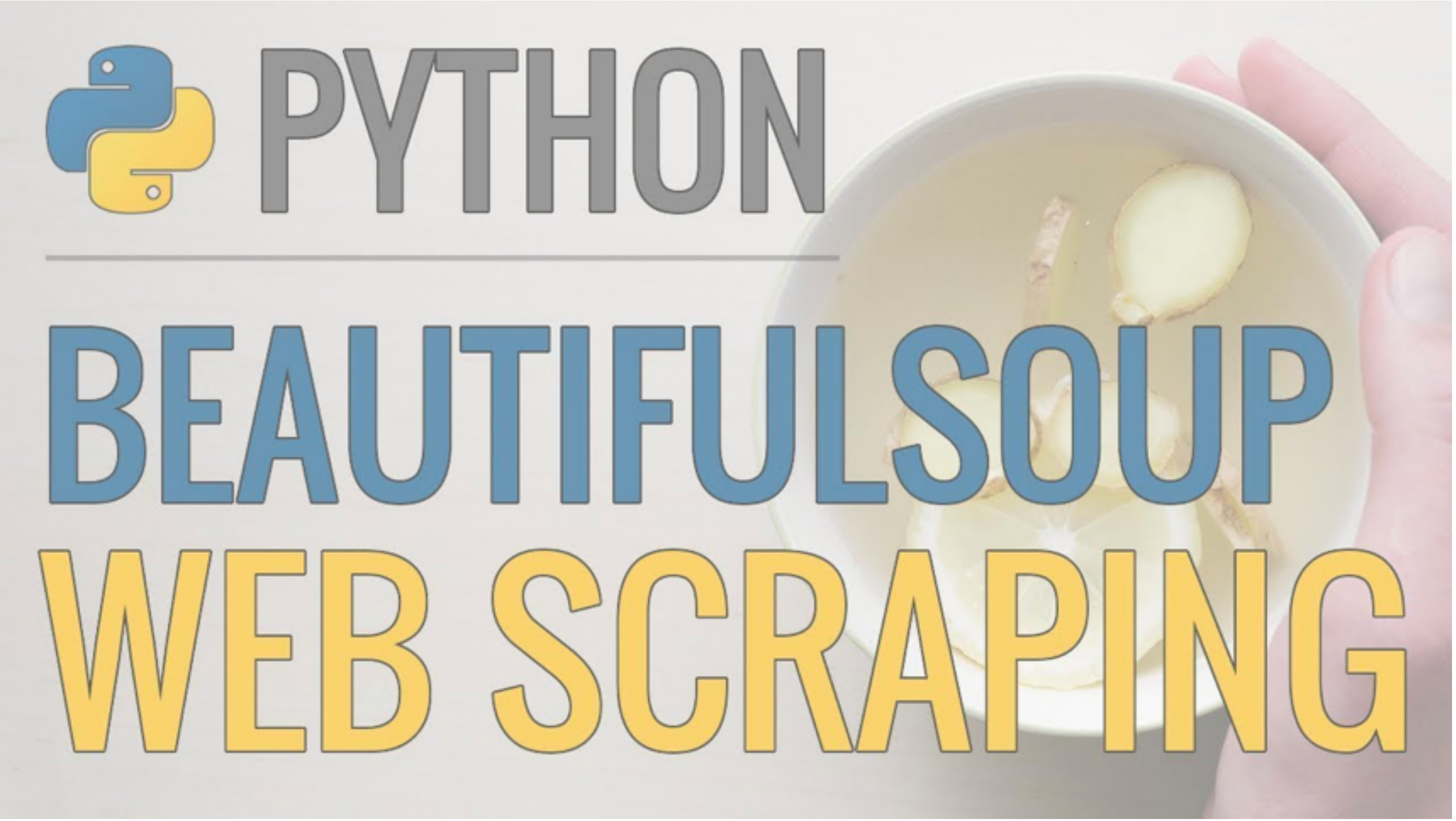
Video files

Email data

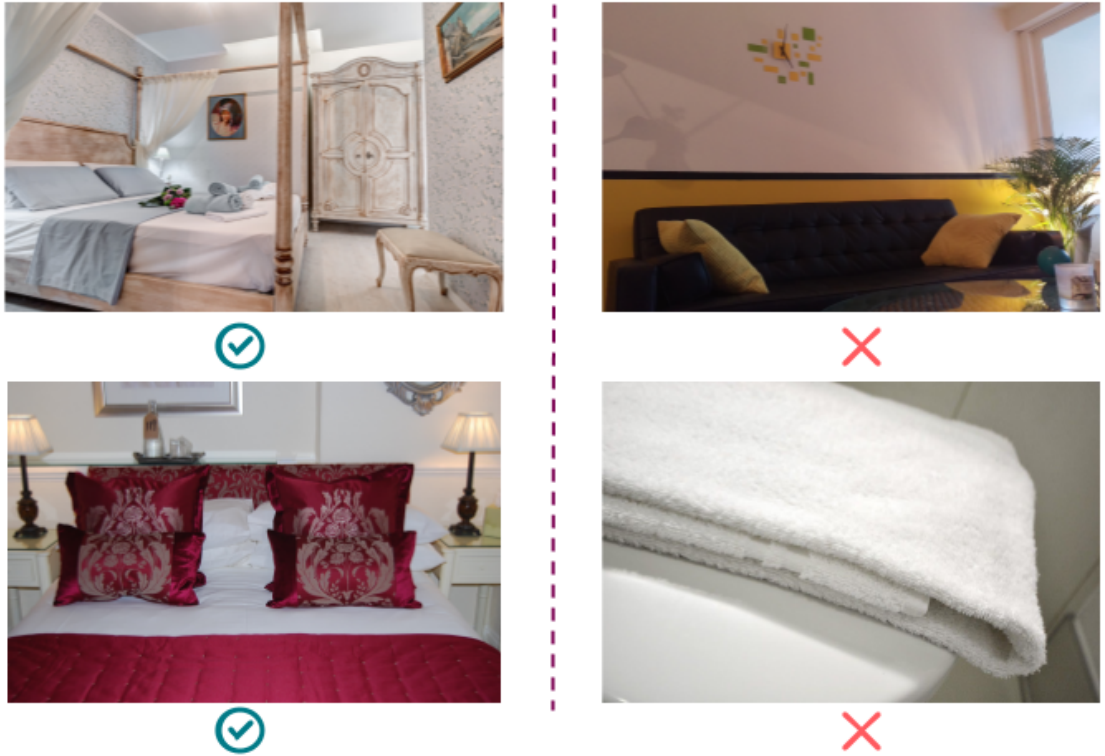
Social media data



Text:
Sentiment Analysis


PYTHON
BEAUTIFULSOUP
WEB SCRAPING

Bedroom Or Not?



“The left two photos were correctly predicted as bedrooms; The right two photos were correctly predicted NOT as bedrooms.”

https://medium.com/airbnb-engineering/categorizing-listing-photos-at-airbnb-f9483f3ab7e3

Tidy Data

"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

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	10,051	9,916	10,265	9,034	9,117	7,659
	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

Primary keynotes

Comma on

Merged cells

Summary of data inside data

Covariate as Subheading

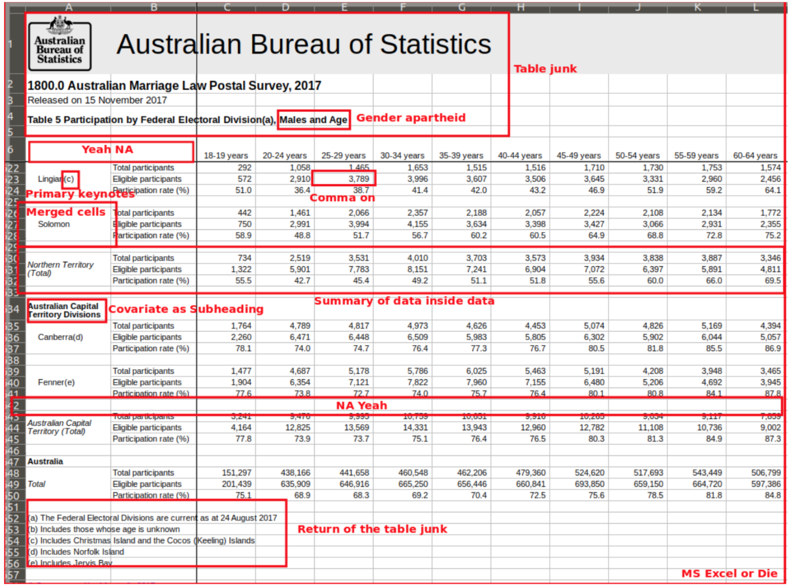
NA Yeah

(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

https://medium.com/@miles.mcbain/tidying-the-a

untidy data



data

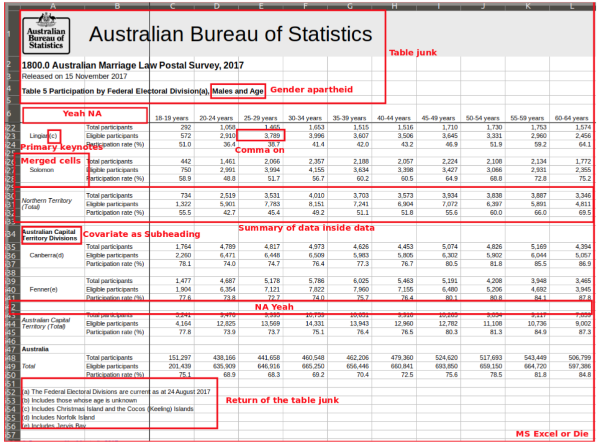
wrangling

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

	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
				SA	76	4342	89.3	3879	3879
				SA	76	3970	90.7	3602	3602
				SA	76	3009	90.3	2716	2716
				SA	76	2156	88.5	1908	1908
				SA	76	1673	85.1	1423	1423



urvey-data-with-r-5d35cea07962

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	Alex	male	Birmingham	AL	Teacher

	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

TIDY data is NOT the same as CLEAN data

Clean Data

Data cleaning is the process of detecting and correcting/ removing data records that are

- incomplete,
- incorrect,
- inaccurate, or
- irrelevant

and often includes making sure missing data is correctly marked and that duplicate records and other errors are removed

Data that is TIDY and CLEAN is ready for use

Tabular Data Time

https://forms.gle/6EKypPu1odgBdbiJ8

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 tables stores data best?

A B C D

Clean

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

Tidy

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

Tidy + Clean

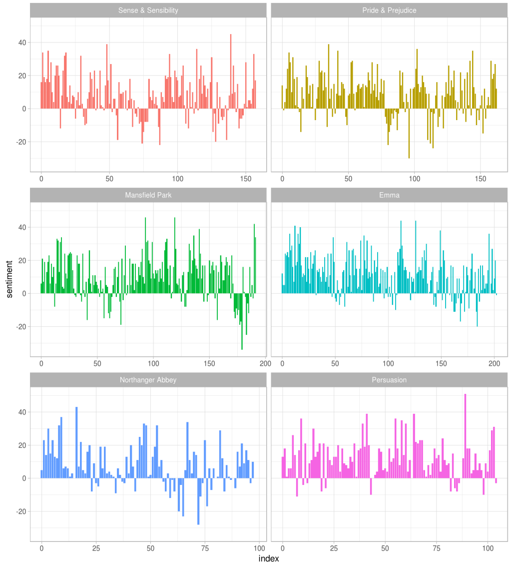
ID	Last	First	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

text

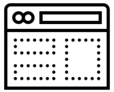
tidy dataset

Word	Novel	Frequency
good	Emma	359
young	Emma	192
friend	Emma	166

results



https://www.tidytextmining.com



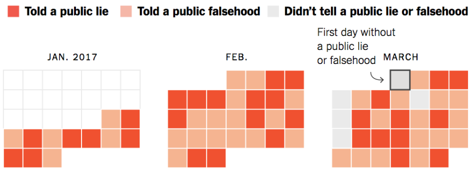
website

JAN. 21 "I wasn't a fan of Iraq. I didn't want to go into Iraq." *(He was for an invasion before he was against it.)* **JAN. 21** "A reporter for Time magazine — and I have been on their cover 14 or 15 times. I think we have the all-time record in the history of Time magazine." *(Trump was on the cover 11 times and Nixon appeared 55 times.)* **JAN. 23** "Between 3 million and 5 million illegal votes caused me to lose the popular vote." *(There's no evidence of illegal voting.)* **JAN. 25** "Now, the audience was the biggest ever. But this crowd was massive. Look how far back it goes. This crowd was massive." *(Official aerial photos show Obama's 2009 inauguration was much more heavily attended.)* **JAN. 25** "Take a look at the Pew reports (which show voter fraud.)" *(The report never mentioned voter fraud.)* **JAN. 25** "You had millions of people that now aren't insured anymore." *(The real number is less than 1 million, according to the Urban Institute.)* **JAN. 25** "So, look, when President Obama was there two weeks ago making a speech, very nice speech. Two people were shot and killed during his speech. You can't have that." *(There were no gun homicide victims in Chicago that day.)* **JAN. 26** "We've taken in tens of thousands of people. We know nothing about them. They can say they vet them. They didn't vet them. They have no papers. How can you vet somebody when you don't know anything about them and you have no papers? How do you vet them? You can't." *(Vetting lasts up to two years.)* **JAN. 26** "I cut off hundreds of millions of dollars off one particular plane, hundreds of millions of dollars in a short period of time. It wasn't like I spent, like, weeks, hours, less than hours, and many, many hundreds of millions of dollars. And the plane's going to be better." *(Most of the cuts were already planned.)* **JAN. 28** "The coverage about me in the @nytimes and the @washingtonpost has been so false and angry that the Times actually apologized to its dwindling subscribers and readers." *(It never apologized.)* **JAN. 29** "The Cuban-Americans, I got 84 percent of that vote." *(There is no support for this.)* **JAN. 30** "Only 109 people out of 325,000 were detained and held for questioning. Big problems at airports were caused by Delta computer outage." *(At least 746 people were detained and processed, and the Delta outage happened two days later.)* **FEB. 3** "Professional anarchists, thugs and paid protesters are proving the point of the millions of people who voted to MAKE AMERICA GREAT AGAIN!" *(There is no evidence of paid protesters.)* **FEB. 4** "After being forced to apologize for its bad and inaccurate coverage of me after winning the election, the FAKE NEWS @nytimes is still lost!" *(It never apologized.)* **FEB. 5** "We had 109 people out of hundreds of thousands of travelers and all we did was vet those people very, very carefully." *(About 60,000 people were affected.)* **FEB. 6** "I have already saved more than $700 million when I got involved in the negotiation on the F-35." *(Much of the price drop was projected before Trump took office.)* **FEB. 6** "It's gotten to a point where it is not even being reported. And in many cases, the very, very dishonest press doesn't want to report it."

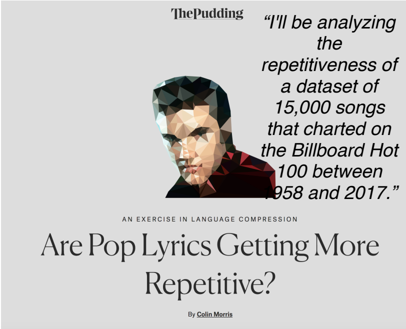
tidy dataset

	date	lie	explanation	url
0	Jan. 21, 2017	I wasn't a fan of Iraq. I didn't want to go in...	He was for an invasion before he was against it.	https://www.buzzfeed.com/andrewkaczynski/in-20...
1	Jan. 21, 2017	A reporter for Time magazine — and I have been...	Trump was on the cover 11 times and Nixon appe...	http://nation.time.com/2013/11/06/10-things-yo...
2	Jan. 23, 2017	Between 3 million and 5 million illegal votes ...	There's no evidence of illegal voting.	https://www.nytimes.com/2017/01/23/us/politics...
3	Jan. 25, 2017	Now, the audience was the biggest ever. But th...	Official aerial photos show Obama's 2009 inaug...	https://www.nytimes.com/2017/01/21/us/politics...
4	Jan. 25, 2017	Take a look at the Pew reports (which show vot...	The report never mentioned voter fraud.	https://www.nytimes.com/2017/01/24/us/politics...

results



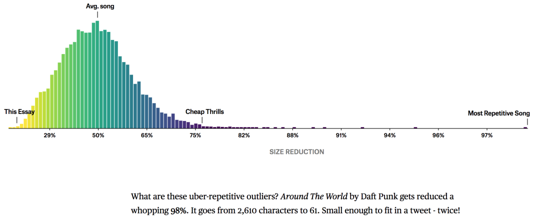
text (lyrics)



tidy dataset

song	Artist	Released	Reduction
Cheap Thrills	Sia	2016	76
Around The World	Daft Punk	1997	98
Everybody Dies	J. Cole	2018	27

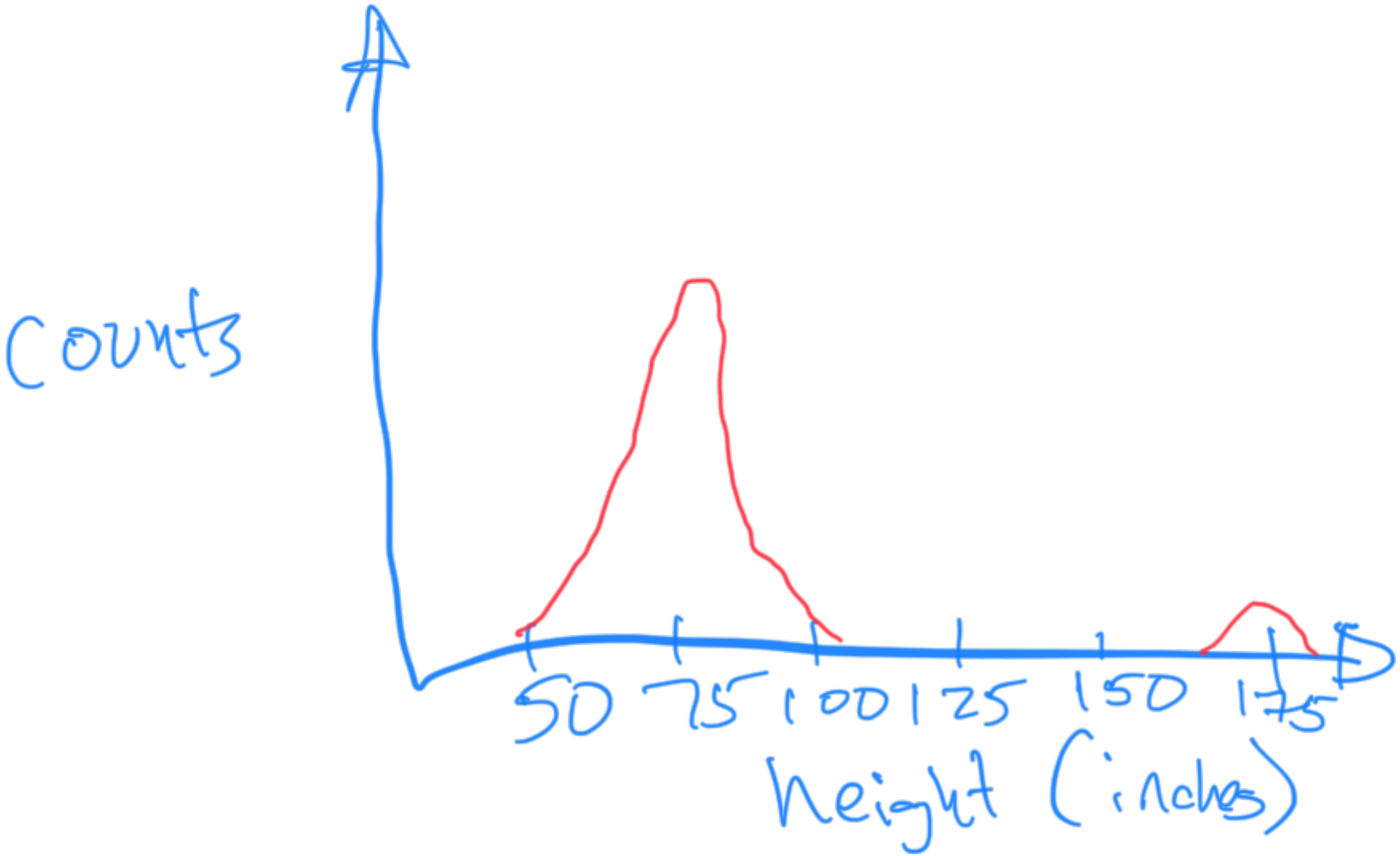
results



https://pudding.cool/2017/05/song-repetition/

Data Intuition

Why do you need intuition?



Fermi Estimation

https://forms.gle/C982naWtU9RvHqAb7

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

A	B	C	D	E
10	100	1000	10,000	100,000

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

Fermi Estimation

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

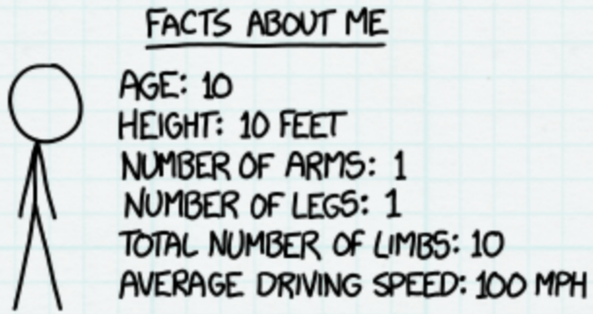
A
YES

B
NO

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.

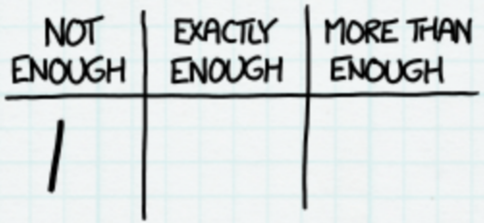
https://what-if.xkcd.com/84/

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:



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

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.

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

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.

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

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.

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

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?

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

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 $1 + \frac{1}{n}$.

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

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.

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

Data Intuition

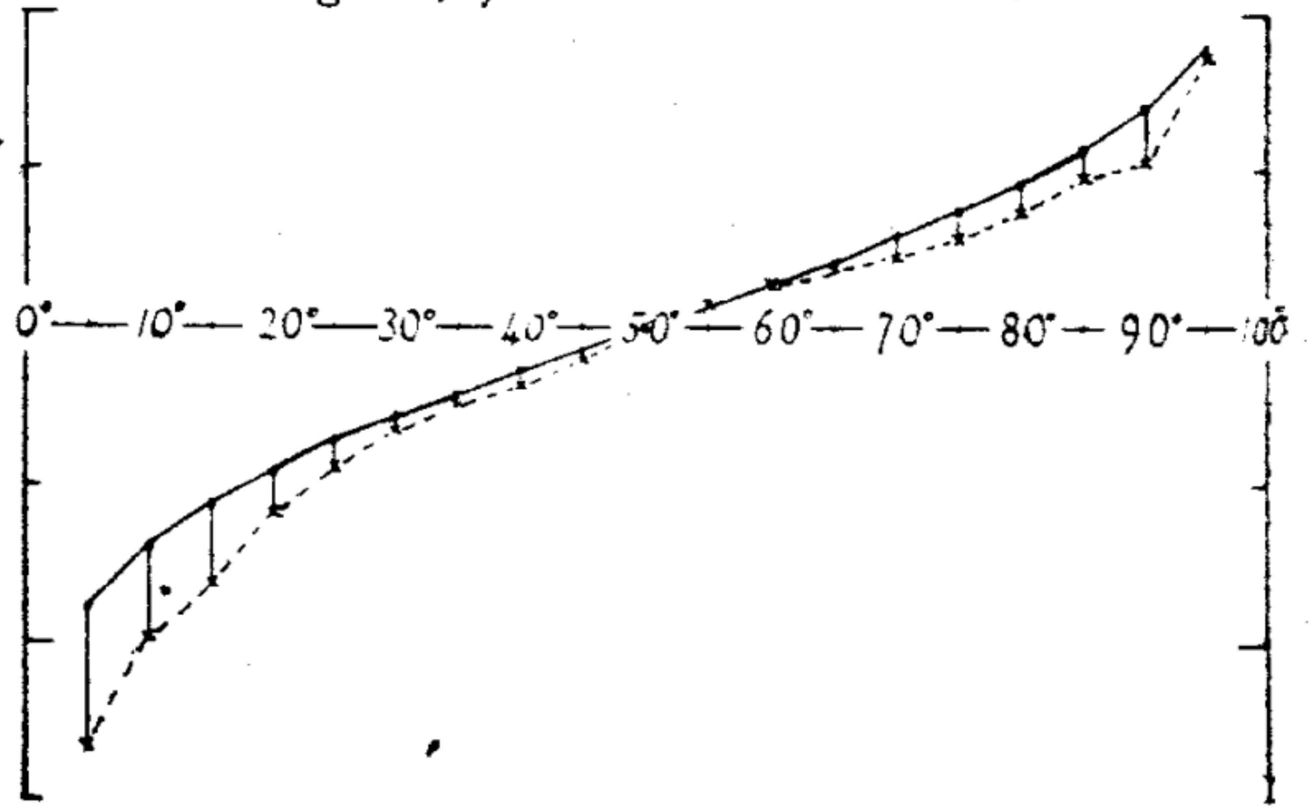
1. Think about your question and your expectations
2. Do some Fermi calculations (back of the envelope calculations)
3. Write code & look at outputs <- think about those outputs
4. Use your gut instinct / background knowledge to guide you
5. Review code & fix bugs

https://forms.gle/CREcpMkYDLYTUp2s6

6 min think
6 min pair
6 share

Other kinds of guessing and intuitions

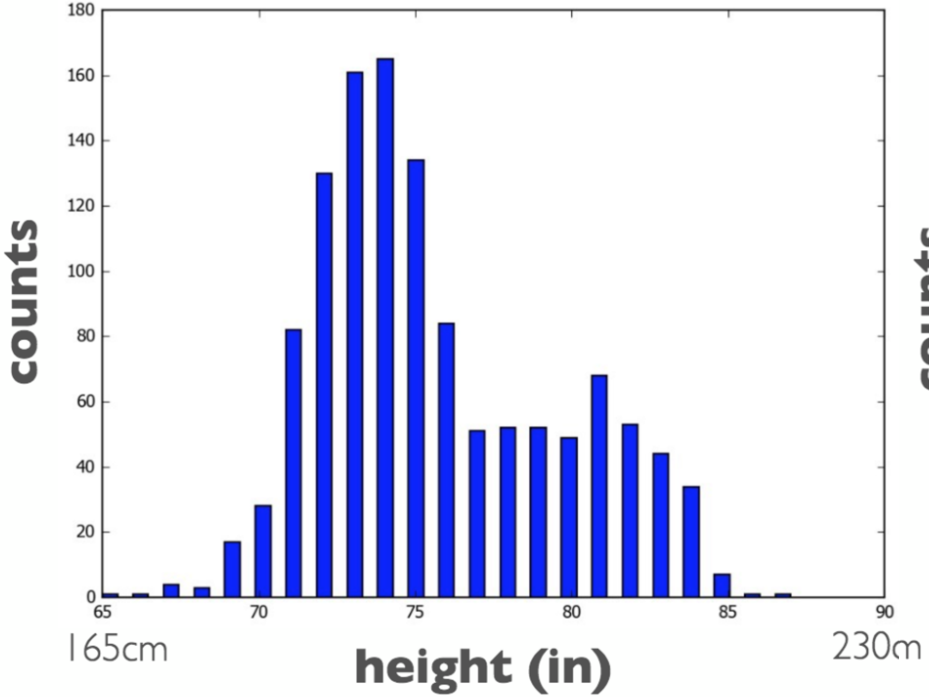
Diagram, from the tabular values.

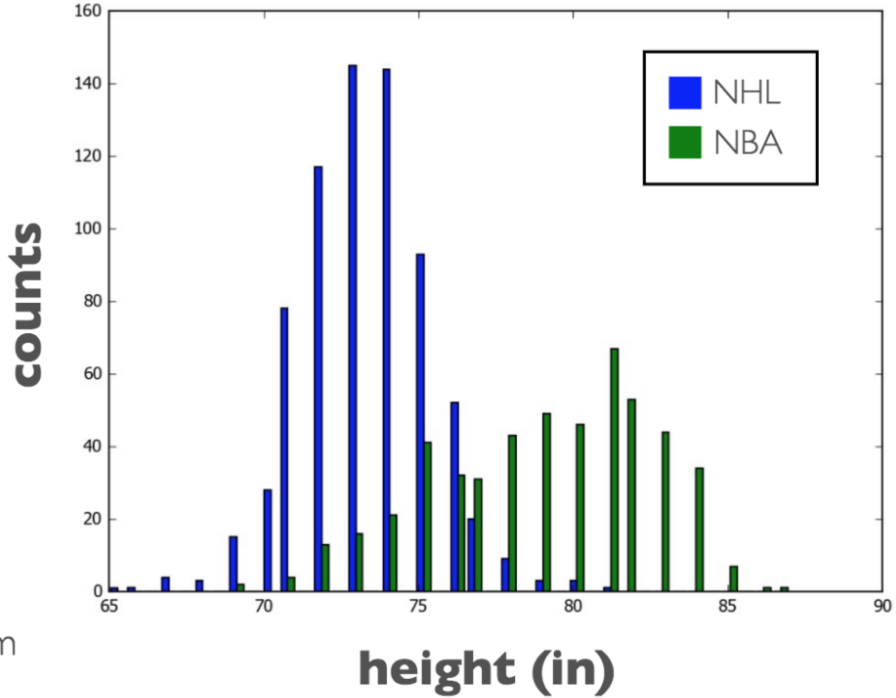


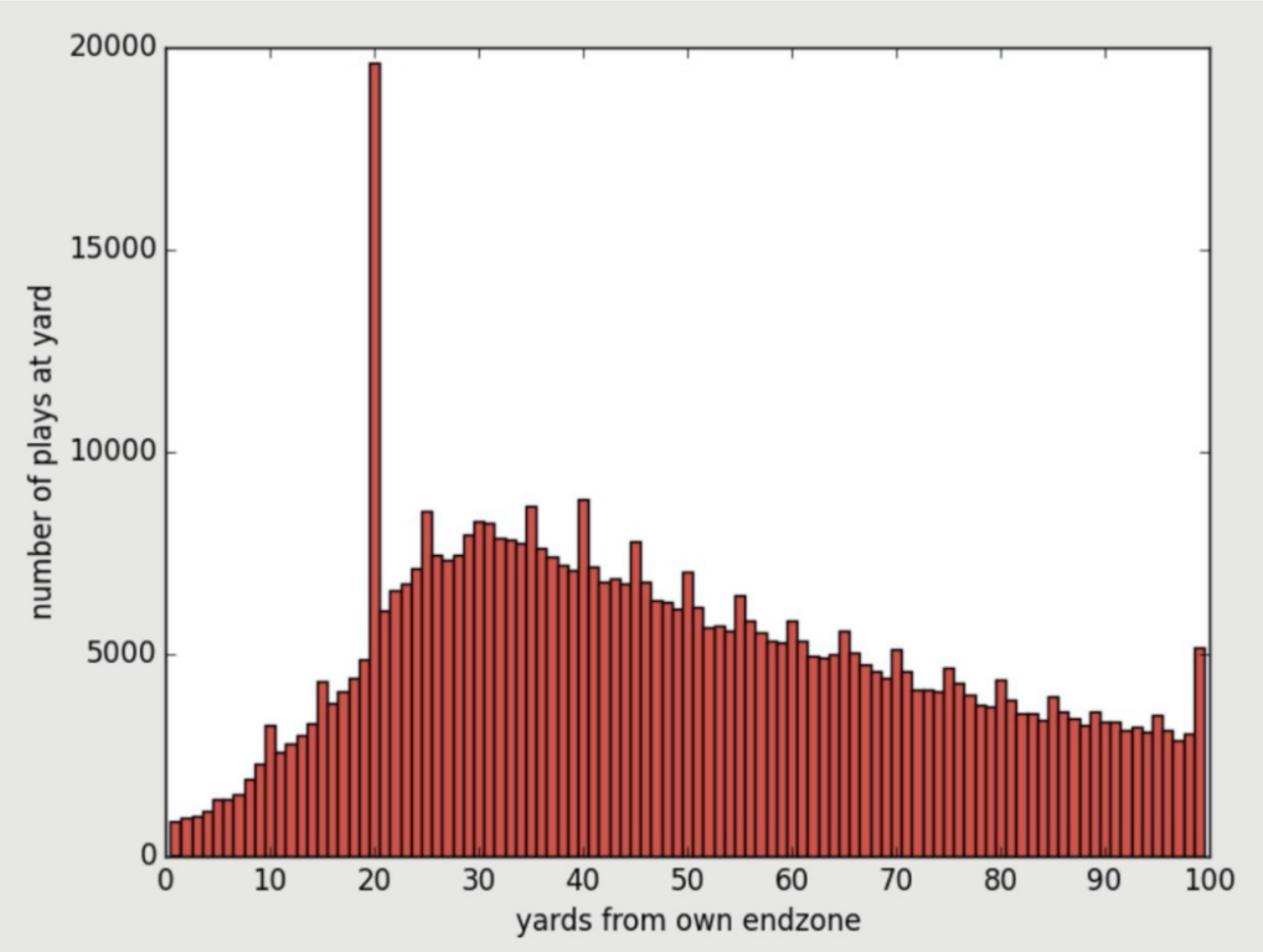
Vox Populi

The Wisdom of the Crowds

- Diversity of opinion: Each person should have private information….even if it's just an eccentric interpretation of the known facts
- Independence: People's opinions aren't determined by the opinions of those around them
- Decentralization: People are able to specialize and draw on local knowledge
- Aggregation: Some mechanism exists for turning private judgements into a collective decision


180
160
140
120
100
80
60
40
20
0
counts
65
70
75
80
85
90
165cm
height (in)
230m


160
140
120
100
80
60
40
20
0
counts
NHL
NBA
65
70
75
80
85
90
height (in)


20000
15000
10000
5000
0
number of plays at yard
0
10
20
30
40
50
60
70
80
90
100
yards from own endzone