Interviewing Your Data: Getting Data

Miklós Koren

CODEDTHINKING.
by Koren

Tidy data

Data vs story

Contrary to the aphorism, data is not the plural of anecdote.

- known structure
- known collection method

Benefits of (more) data

- Less subject to selection/recall bias.
- Statistical learning can be applied.

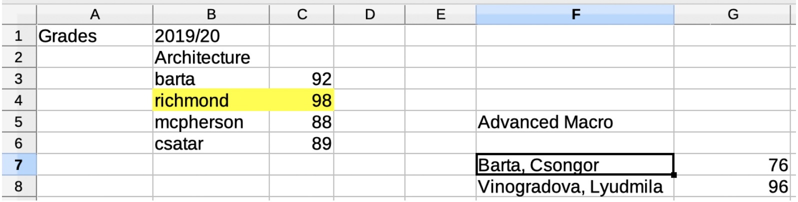
Drawbacks of (more) data

- Harder to work with.
- Harder to interpret/relate to.
- May give false sense of knowledge.

Tidy data

1. Every row is an observation (case, record)
2. Every column is a variable (feature, attribute)
3. Every cell contains a single value

Spreadsheet as a canvas



	A	B	C	D	E	F	G
1	Grades	2019/20					
2		Architecture					
3		barta	92				
4		richmond	98				
5		mcpherson	88			Advanced Macro	
6		csatar	89				
7						Barta, Csongor	76
8						Vinogradova, Lyudmila	96

Same content in tidy form

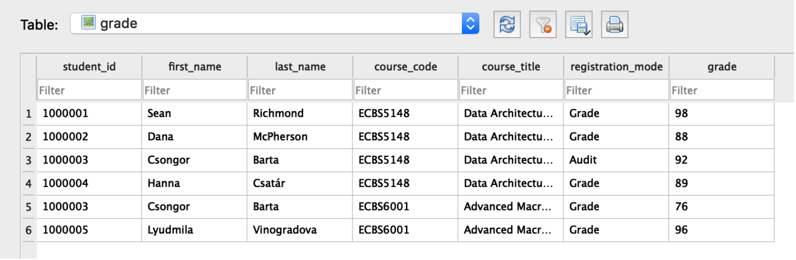


Table: grade

	student_id	first_name	last_name	course_code	course_title	registration_mode	grade
	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1000001	Sean	Richmond	ECBS5148	Data Architectu...	Grade	98
2	1000002	Dana	McPherson	ECBS5148	Data Architectu...	Grade	88
3	1000003	Csongor	Barta	ECBS5148	Data Architectu...	Audit	92
4	1000004	Hanna	Csatár	ECBS5148	Data Architectu...	Grade	89
5	1000003	Csongor	Barta	ECBS6001	Advanced Macr...	Grade	76
6	1000005	Lyudmila	Vinogradova	ECBS6001	Advanced Macr...	Grade	96

Benefits

- Machine readable
- Can select columns (=variables)
- Can filter rows (=observations)
- Can sort rows
- Can join rows

Spreadsheets

Learn to love your spreadsheet editor

- Beware of Excel!
- Good alternatives: Libre Office, Open Office, Google Sheets.

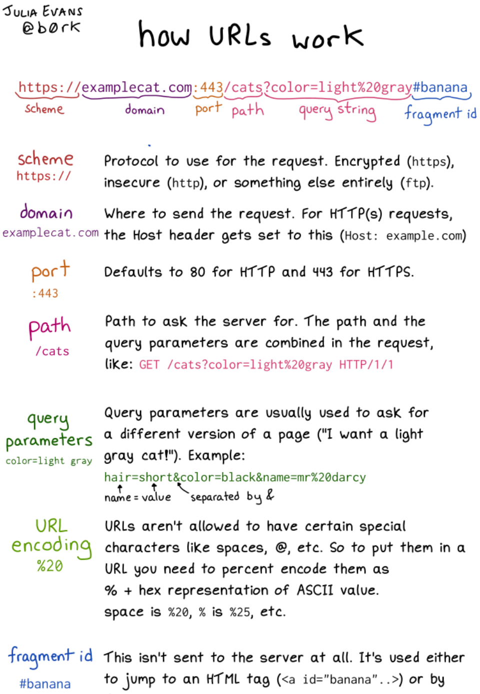
Useful steps

- filter
- sort
- vlookup

Exercise

Getting data

What's in a URL?



JULIA EVANS
@b0rk

how URLs work

`https://examplecat.com:443/cats?color=light%20gray#banana`

scheme `https://` — Protocol to use for the request. Encrypted (`https`), insecure (`http`), or something else entirely (`ftp`).

domain `examplecat.com` — Where to send the request. For HTTP(s) requests, the Host header gets set to this (`Host: example.com`)

port `:443` — Defaults to 80 for HTTP and 443 for HTTPS.

path `/cats` — Path to ask the server for. The path and the query parameters are combined in the request, like: `GET /cats?color=light%20gray HTTP/1/1`

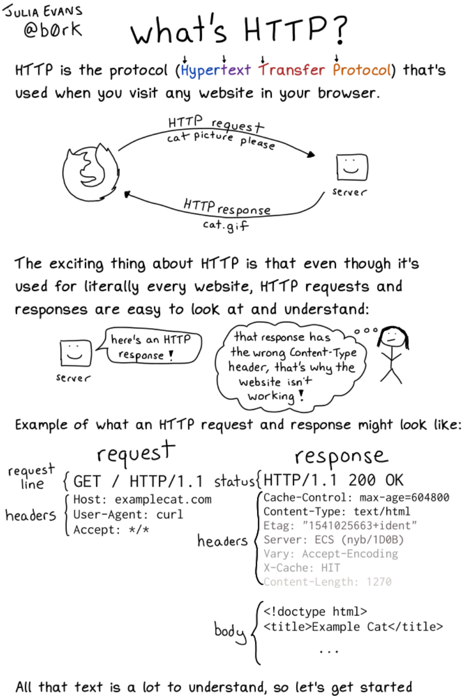
query parameters `color=light gray` — Query parameters are usually used to ask for a different version of a page ("I want a light gray cat!"). Example:

`hair=short&color=black&name=mr%20darcy`

URL encoding `%20` — URLs aren't allowed to have certain special characters like spaces, @, etc. So to put them in a URL you need to percent encode them as % + hex representation of ASCII value. space is `%20`, `%` is `%25`, etc.

fragment id `#banana` — This isn't sent to the server at all. It's used either to jump to an HTML tag (`<a id="banana"..>`) or by

What’s HTTP?



JULIA EVANS
@b0rk

what's HTTP?

HTTP is the protocol (Hypertext Transfer Protocol) that's used when you visit any website in your browser.

HTTP request
cat picture please
server
HTTP response
cat.gif

The exciting thing about HTTP is that even though it's used for literally every website, HTTP requests and responses are easy to look at and understand:

here's an HTTP response!
server
that response has the wrong Content-Type header, that's why the website isn't working!

Example of what an HTTP request and response might look like:

request

request line
```
GET / HTTP/1.1
```
headers
```
Host: examplecat.com
User-Agent: curl
Accept: */*
```

response

status
```
HTTP/1.1 200 OK
```
headers
```
Cache-Control: max-age=604800
Content-Type: text/html
Etag: "1541025663+ident"
Server: ECS (nyb/1D0B)
Vary: Accept-Encoding
X-Cache: HIT
Content-Length: 1270
```
body
```
<!doctype html>
<title>Example Cat</title>
       ...
```

All that text is a lot to understand, so let's get started

API, CSV, XML and JSON

The world of data is full of acronyms.

- API: Application Programming Interface, the language in which machines talk to one another. Useful for automating data gathering and updating.
- CSV: Comma Separated Values, a plain text format for data tables. Everything can read it and write it (beware of Excel).
- XML: Extensible Markup Language, a structured document format to store hierarchical data. Very widely used, but not human friendly.
- JSON: JavaScript Object Notation, the de facto web standard for sharing structured data. Similar to XML, but much more legible.

Scraping

scraping = crawling + parsing

Four steps of a scraping project

1. Recon
2. Crawl
3. Parse
4. Store

Recon

1. Locate the interesting documents and tables
2. Note the structure of URLs and tables
3. Explore `robots.txt` and terms of use
4. Explore robot protection

Crawl

1. Download the pages you need.
2. Verify that you have the correct number of pages.

Parse

1. Find and extract the information within the HTML structure.
2. Verify that you have everything you need. (Save link to original!)

Crawling and parsing often done together in scraping apps.

Exercise