Class 7: Data Science Foundations

MA5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.github.io

11 December 2023

Outline of Today's Class

Some Housekeeping ...

Spring Term Learning Outcomes

1. Collect web data independently
   - Building Web Scrapers via R
2. Carry out text analysis via R

MAST5953: Web Scraping & Text Mining

Spring Term Structure

1. Class 7 (today): Data Science/Webscraping Primer
2. Class 8 (18/01): Web Scraping and Regular Expressions
3. Class 9 (18/01): Scraping Speech Data
4. Class 10 (25/01): Text Mining I: Pre-Processing and the Document-Term Matrix
5. Class 11 (25/01): Text Mining II: Topic Models

Spring Term Assessment

- 1,000 words Web Scraping / Text Mining Task
  - Due: 18th March 2024 - **4pm!**
  - For issues with submission, and to request extensions/mitigation contact CEMS Student Support: cemssupport@kent.ac.uk

Spring Term Assessment

The Assessment Brief

Scrape the speeches of 2 politicians of your choosing, **on the opposite spectrum ideologically**. You will then have to answer the **question: how do these politicians differ** in their speeches? You will answer this question by interpreting (a) text descriptives; (b) results from a topic model analysis. The 1,000 words report will include (1) a section **explaining what web-scraping is, what method/process was undertaken to scrape the speeches**; (2) a section **presenting the text descriptive statistics and graphs and interpretation of those**; (3) a section where the **topic modelling method is briefly explained and where the results from the comparison are presented** (with visualisations and/or numerical summaries).

Spring Term Assessment

What I will be assessing

- Technical, professional, concise explanations of web-scraping as a method of data collection and what specific web-scraping method was used for the report.
- A technical, professional, concise presentation of text descriptives, covering all relevant descriptive statistics and graphs with clear and correct interpretations of those.
- Technical, professional, concise explanations of topic modelling as a method for text analysis and clear and correct presentation/interpretation of the comparison results.
- Good formatting and well presented R code and/or Appendices (if any added).

Your Feedback ...

- Mid-Survey Evaluation

R Revision + Key Definitions

Course Requirements: R

Particularly for the Spring Term Sessions!

- Prior knowledge of `R` is a must, make sure you understand the language basics (packages, objects/vectors, core functions and vector + data management operations), and that you know how to trouble-shoot errors and install packages
  - Great `R` Intro Resources:
    - Adler, Joseph. 2009. *R in a Nutshell. A Desktop Quick Reference*. O'Reilly
    - Teetor, Paul. 2011. *R Cookbook*. O'Reilly.
    - I recommend the following website for revisions: `https://stats.idre.ucla.edu/r/`

Course Requirements: R

- Make sure you have installed `R` and `RStudio` and everything is up-to-date.
  - `R`:
    - Type `rversions::r_release()` in R to check
    - Follow the instructions to download R here: R Installer
  - `RStudio`:
    - Updates needed: Go to «Help» - «Check for Updates» - & follow instructions in the pop-up
    - To download RStudio: RStudio Installer

R debugging

Examples of some common errors

- If it says: there is no package called 'xxx' you just need to type:
  - install.packages("xxx")
- If it says: could not find function 'xxx', just call the library by typing:
  - `require(xxx)` or `library(xxx)`
- If it says: cannot open the connection, it means you have not specified the correct working directory (i.e. folder) where the file is located
  - `setwd()`
- Great Resources to Trouble-Shoot Errors:
  - StackOverflow
  - Stack Exchange
  - **just Google it, it's not cheating** :)

Preach!

John B. Holbein
@JohnHolbein1

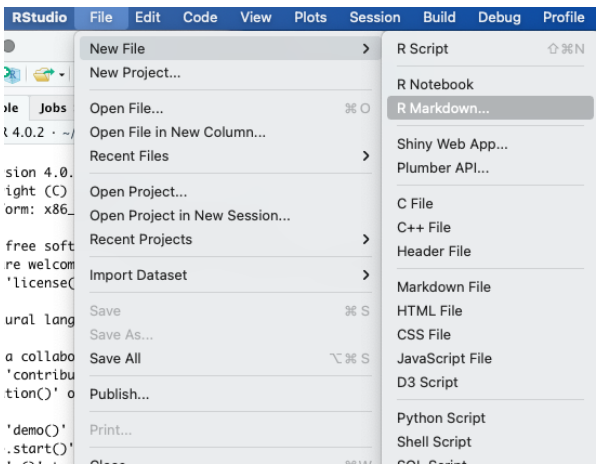
Students often don't realize that being good at coding really just means being good at Googling.

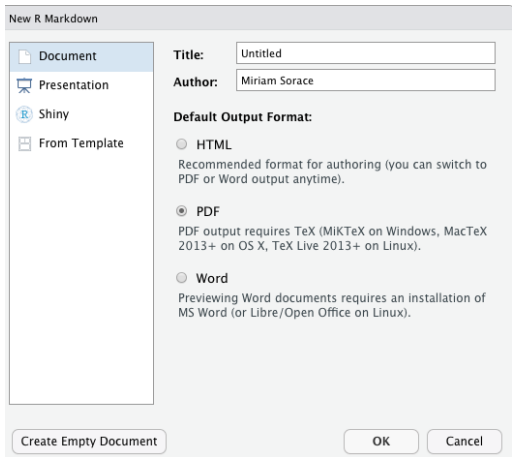
3:03 PM · Jan 11, 2021 · Twitter Web App

32 Retweets **18** Quote Tweets **554** Likes

Creating an RMarkdown File

- From scratch:





- Opening an existing file/template: «File» - «Open File» - Navigate to the folder where you saved the .Rmd file & select the file.
- Revise the RMarkdown Cheat Sheet

Problems with Knitting?

- If all troubleshooting fails, try to select the 'Word' or 'HTML' options instead of the pdf option when creating your RMarkdown
  - E.g. instead of `output: pdf_document`, use `output: html_document`
  - You can then convert to a pdf later (online pages or via 'Save as' in your laptop)!

Key Terms: What is Data?

- Data = coded information for statistical processing
  - letters from an historical archive
  - legislative speeches
  - survey answers
  - tweets
  - Instagram pictures
- Data can be *primary* (collected directly by the researcher - e.g. original survey) or *secondary* (taken from an existing source - e.g. web database)

Key Terms: Data Science

- Collecting, organising and analysing 'big data'
- Advent of Web means large amounts of data are online
- Programming skills (R features prominently) are a must

Key Terms: Web Scraping

- Building a computer program (*scraper*) to grab specific content from webpages and convert it into usable datasets
- $\neq$ Spidering/Web Crawling: grabbing entire webpage and links in an unstructured way

Disclaimer

you will quickly learn that there is no 'universal' web scraper: web pages are never the same and the same web page can change over time! Finding the perfect scraper to your specific data collection needs requires *customised* programs!

Web Scraping: When Should I Bother?

- data collection task is repeated (e.g. database update)
- data collection task is complex
  - large data
  - expensive (time or money) to do manually
- Web scraping is less error-prone, more time-efficient, and fully reproducible!

Web Scraping Example

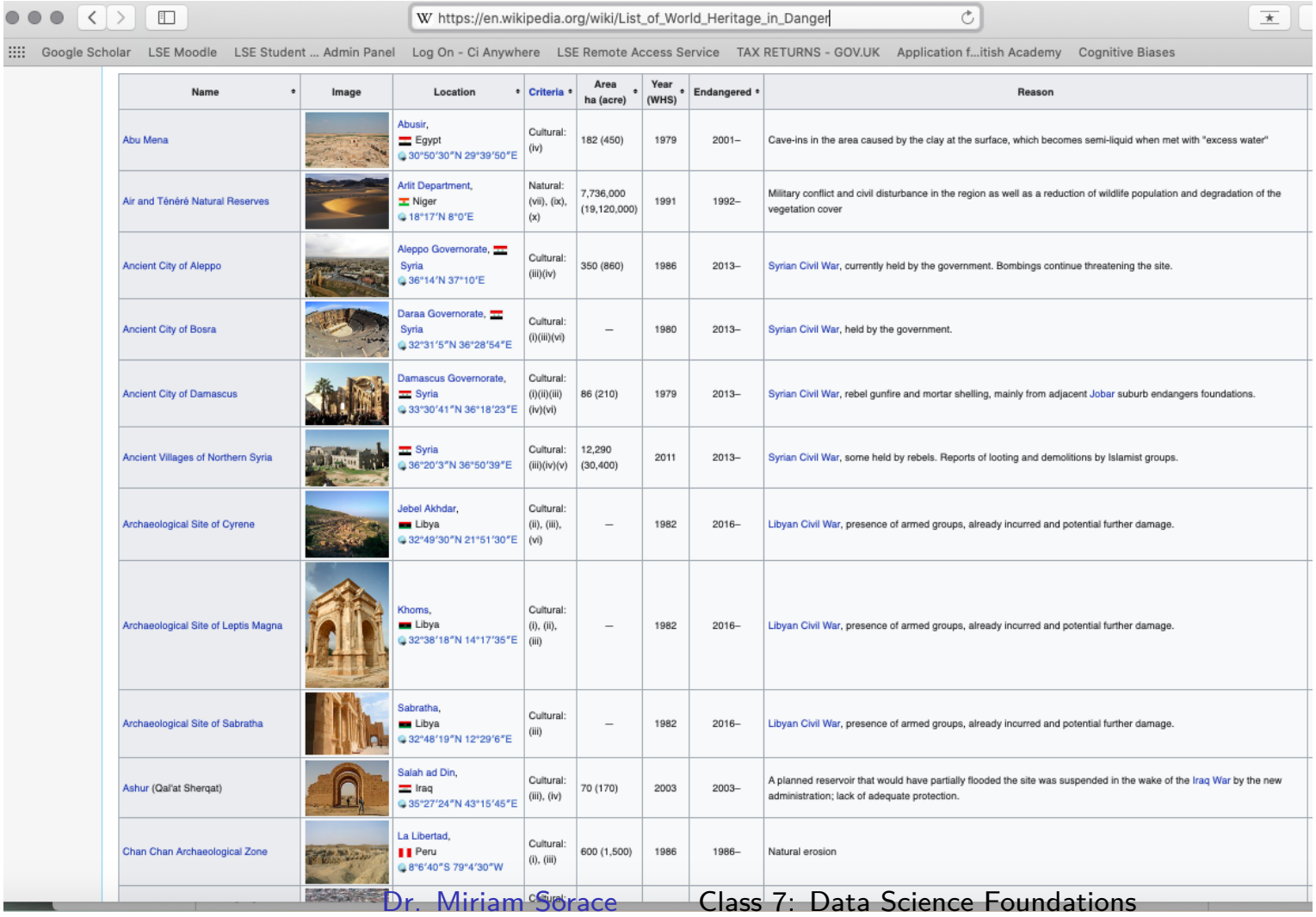
How Facebook Started

Check out this short video from the movie:
The Social Media - the 'hacking/data scraping' scene

Web Scraping Example

Heritage Sites in Danger

Which sites are threatened and where are they located?



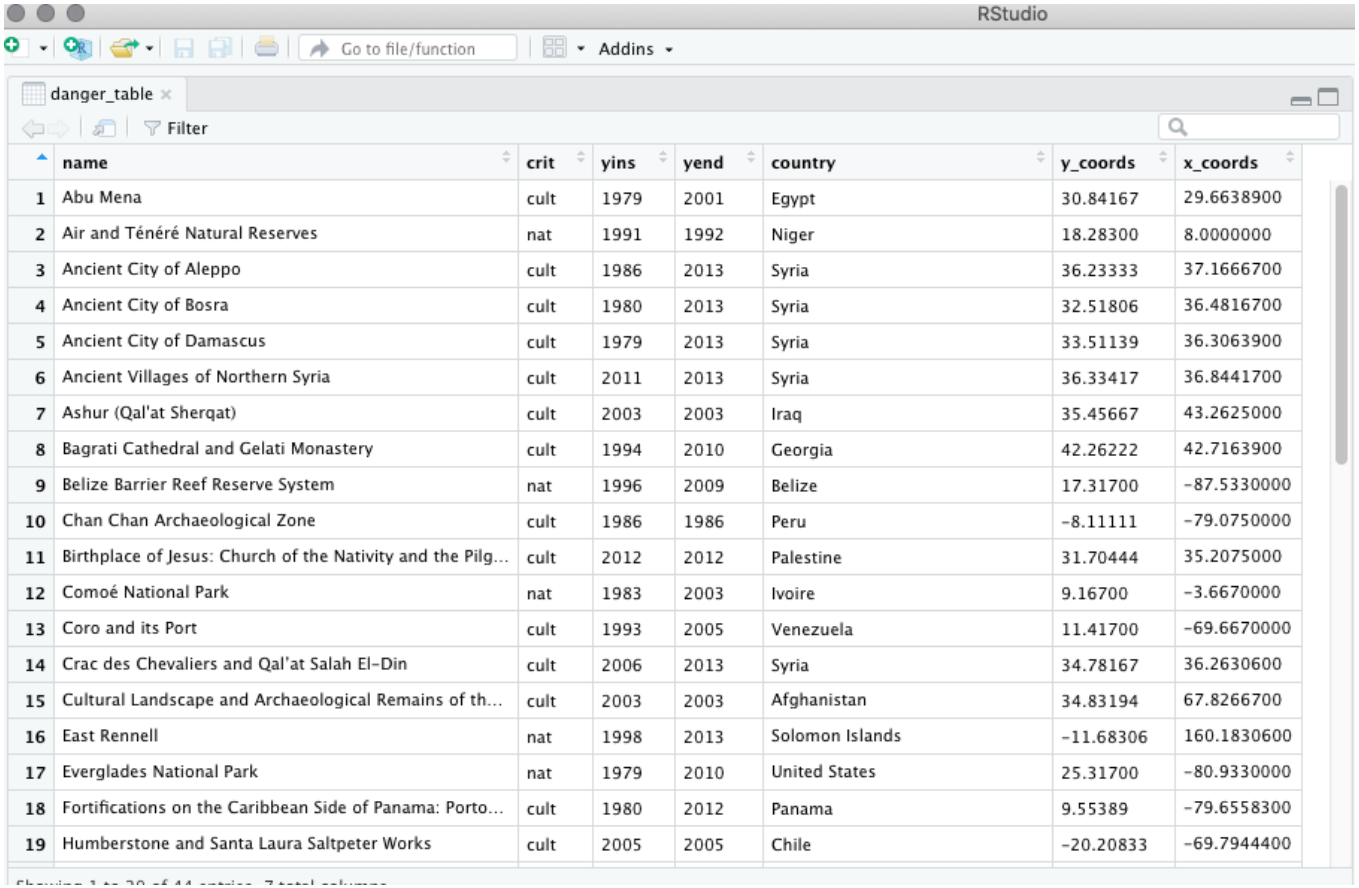
https://en.wikipedia.org/wiki/List_of_World_Heritage_in_Danger

Google Scholar | LSE Moodle | LSE Student ... Admin Panel | Log On - Ci Anywhere | LSE Remote Access Service | TAX RETURNS - GOV.UK | Application f...itish Academy | Cognitive Biases

Name	Image	Location	Criteria	Area ha (acre)	Year (WHS)	Endangered	Reason
Abu Mena		Abusir, Egypt 30°50′30″N 29°39′50″E	Cultural: (iv)	182 (450)	1979	2001–	Cave-ins in the area caused by the clay at the surface, which becomes semi-liquid when met with "excess water"
Air and Ténéré Natural Reserves		Arlit Department, Niger 18°17′N 8°0′E	Natural: (vii), (ix), (x)	7,736,000 (19,120,000)	1991	1992–	Military conflict and civil disturbance in the region as well as a reduction of wildlife population and degradation of the vegetation cover
Ancient City of Aleppo		Aleppo Governorate, Syria 36°14′N 37°10′E	Cultural: (iii)(iv)	350 (860)	1986	2013–	Syrian Civil War, currently held by the government. Bombings continue threatening the site.
Ancient City of Bosra		Daraa Governorate, Syria 32°31′5″N 36°28′54″E	Cultural: (i)(iii)(vi)	–	1980	2013–	Syrian Civil War, held by the government.
Ancient City of Damascus		Damascus Governorate, Syria 33°30′41″N 36°18′23″E	Cultural: (i)(ii)(iii) (iv)(vi)	86 (210)	1979	2013–	Syrian Civil War, rebel gunfire and mortar shelling, mainly from adjacent Jobar suburb endangers foundations.
Ancient Villages of Northern Syria		Syria 36°20′3″N 36°50′39″E	Cultural: (iii)(iv)(v)	12,290 (30,400)	2011	2013–	Syrian Civil War, some held by rebels. Reports of looting and demolitions by Islamist groups.
Archaeological Site of Cyrene		Jebel Akhdar, Libya 32°49′30″N 21°51′30″E	Cultural: (ii), (iii), (vi)	–	1982	2016–	Libyan Civil War, presence of armed groups, already incurred and potential further damage.
Archaeological Site of Leptis Magna		Khoms, Libya 32°38′18″N 14°17′35″E	Cultural: (i), (ii), (iii)	–	1982	2016–	Libyan Civil War, presence of armed groups, already incurred and potential further damage.
Archaeological Site of Sabratha		Sabratha, Libya 32°48′19″N 12°29′6″E	Cultural: (iii)	–	1982	2016–	Libyan Civil War, presence of armed groups, already incurred and potential further damage.
Ashur (Qal'at Sherqat)		Salah ad Din, Iraq 35°27′24″N 43°15′45″E	Cultural: (iii), (iv)	70 (170)	2003	2003–	A planned reservoir that would have partially flooded the site was suspended in the wake of the Iraq War by the new administration; lack of adequate protection.
Chan Chan Archaeological Zone		La Libertad, Peru 8°6′40″S 79°4′30″W	Cultural: (i), (iii)	600 (1,500)	1986	1986–	Natural erosion

Web Scraping Example

Heritage Sites in Danger - Example Code in GitHub page



RStudio

Go to file/function | Addins

danger_table

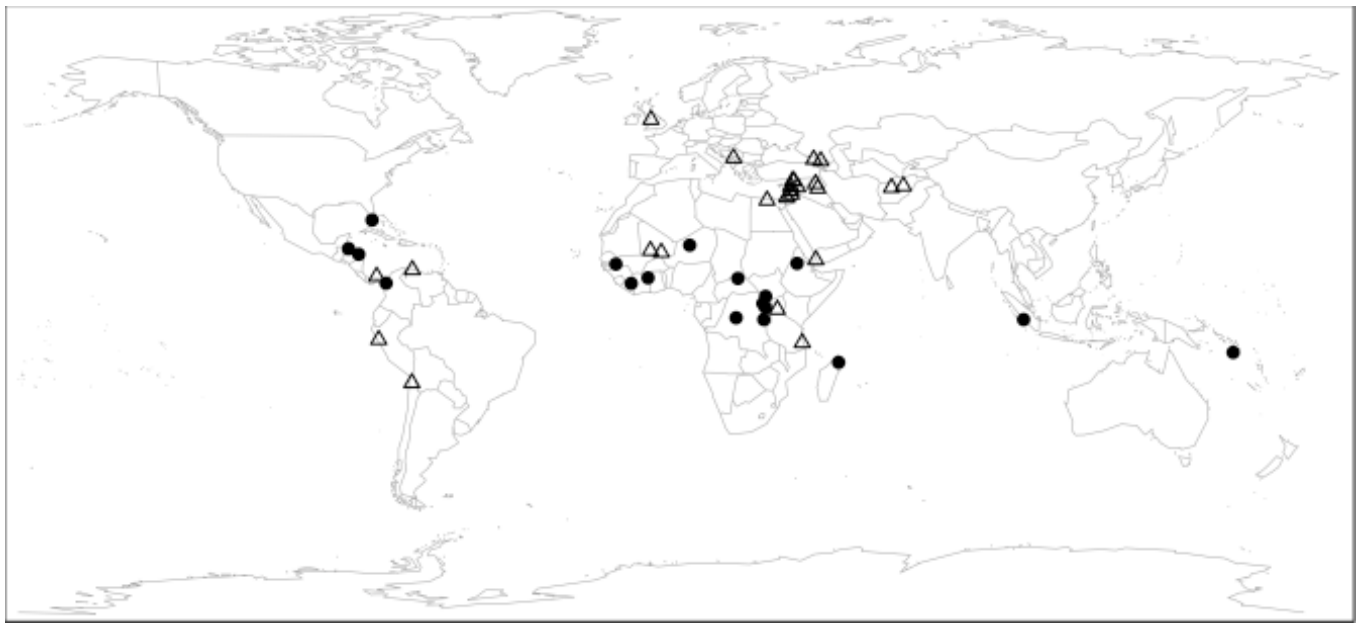
Filter

	name	crit	yins	yend	country	y_coords	x_coords
1	Abu Mena	cult	1979	2001	Egypt	30.84167	29.6638900
2	Air and Ténéré Natural Reserves	nat	1991	1992	Niger	18.28300	8.0000000
3	Ancient City of Aleppo	cult	1986	2013	Syria	36.23333	37.1666700
4	Ancient City of Bosra	cult	1980	2013	Syria	32.51806	36.4816700
5	Ancient City of Damascus	cult	1979	2013	Syria	33.51139	36.3063900
6	Ancient Villages of Northern Syria	cult	2011	2013	Syria	36.33417	36.8441700
7	Ashur (Qal'at Sherqat)	cult	2003	2003	Iraq	35.45667	43.2625000
8	Bagrati Cathedral and Gelati Monastery	cult	1994	2010	Georgia	42.26222	42.7163900
9	Belize Barrier Reef Reserve System	nat	1996	2009	Belize	17.31700	-87.5330000
10	Chan Chan Archaeological Zone	cult	1986	1986	Peru	-8.11111	-79.0750000
11	Birthplace of Jesus: Church of the Nativity and the Pilg...	cult	2012	2012	Palestine	31.70444	35.2075000
12	Comoé National Park	nat	1983	2003	Ivoire	9.16700	-3.6670000
13	Coro and its Port	cult	1993	2005	Venezuela	11.41700	-69.6670000
14	Crac des Chevaliers and Qal'at Salah El-Din	cult	2006	2013	Syria	34.78167	36.2630600
15	Cultural Landscape and Archaeological Remains of th...	cult	2003	2003	Afghanistan	34.83194	67.8266700
16	East Rennell	nat	1998	2013	Solomon Islands	-11.68306	160.1830600
17	Everglades National Park	nat	1979	2010	United States	25.31700	-80.9330000
18	Fortifications on the Caribbean Side of Panama: Porto...	cult	1980	2012	Panama	9.55389	-79.6558300
19	Humberstone and Santa Laura Saltpeter Works	cult	2005	2005	Chile	-20.20833	-69.7944400

From: Munzert, S., Rubba, C., Meißner, P., Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley Sons.

Web Scraping Example

Heritage Sites in Danger



From: Munzert, S., Rubba, C., Meißner, P., Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley Sons.

Key Terms: Text Mining

- Automatic categorization of text data
- Usually done on the basis of document word frequency similarity
- Useful when dealing with textual 'big data', to impose a structure

Data Principles and Ethics

Data Collection Principles

Data Quality

- Is the data suited to answer my research question? Theory before data!
- Is the data source accurate? Cross-validate!
  - Wikipedia is not a qualitative gold standard ...

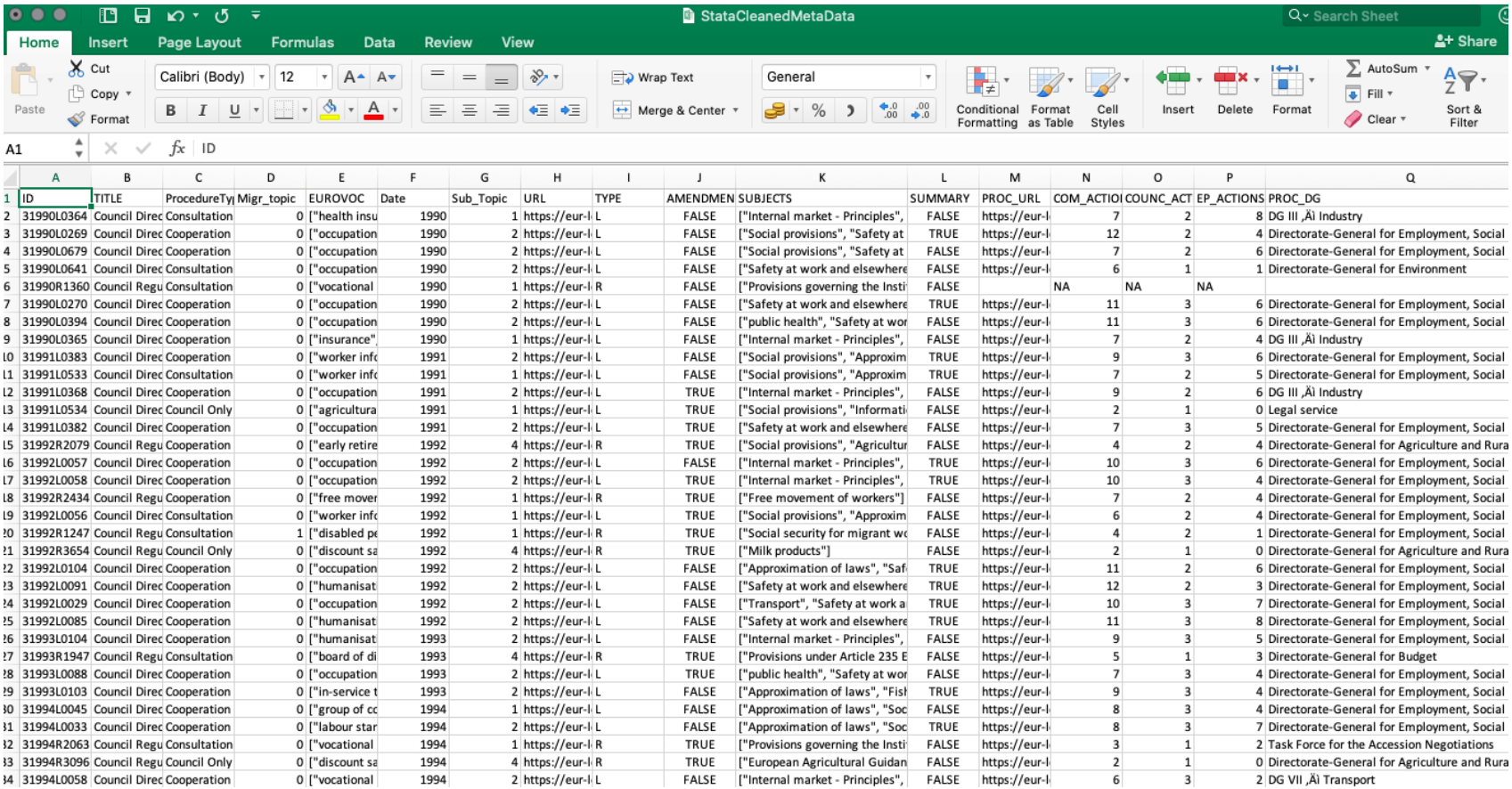
Data Collection Principles

Data Cleaning

- Is the data complete?
- Is the data consistently measured?
  - Use consistent variable names
  - Use consistent file names
- Data should be organised as a single rectangle
  - units in row, variables in columns
  - first row should contain variable names
  - in naming, avoid spaces and special characters. Use hyphens/underscores
  - only 1 piece of observation per cell
  - avoid numbering missing values. Use blank space, "NA", or "."

Data Collection Principles

Example Dataset



StataCleanedMetaData

ID	TITLE	ProcedureTy	Migr_topic	EUROVOC	Date	Sub_Topic	URL	TYPE	AMENDMEN	SUBJECTS	SUMMARY	PROC_URL	COM_ACTIO	COUNC_ACT	EP_ACTIONS	PROC_DG
31990L0364	Council Direc	Consultation	0	["health insu	1990	1	https://eur-l	L	FALSE	["Internal market - Principles",	FALSE	https://eur-l	7	2	8	DG III ,Äì Industry
31990L0269	Council Direc	Cooperation	0	["occupation	1990	2	https://eur-l	L	FALSE	["Social provisions", "Safety at	TRUE	https://eur-l	12	2	4	Directorate-General for Employment, Social
31990L0679	Council Direc	Cooperation	0	["occupation	1990	2	https://eur-l	L	FALSE	["Social provisions", "Safety at	FALSE	https://eur-l	7	2	6	Directorate-General for Employment, Social
31990L0641	Council Direc	Consultation	0	["occupation	1990	2	https://eur-l	L	FALSE	["Safety at work and elsewhere	FALSE	https://eur-l	6	1	1	Directorate-General for Environment
31990R1360	Council Regu	Consultation	0	["vocational	1990	1	https://eur-l	R	FALSE	["Provisions governing the Insti	FALSE		NA	NA	NA	
31990L0270	Council Direc	Cooperation	0	["occupation	1990	2	https://eur-l	L	FALSE	["Safety at work and elsewhere	TRUE	https://eur-l	11	3	6	Directorate-General for Employment, Social
31990L0394	Council Direc	Cooperation	0	["occupation	1990	2	https://eur-l	L	FALSE	["public health", "Safety at wor	FALSE	https://eur-l	11	3	6	Directorate-General for Employment, Social
31990L0365	Council Direc	Cooperation	0	["insurance",	1990	1	https://eur-l	L	FALSE	["Internal market - Principles",	FALSE	https://eur-l	7	2	4	DG III ,Äì Industry
31991L0383	Council Direc	Cooperation	0	["worker infc	1991	2	https://eur-l	L	FALSE	["Social provisions", "Approxim	TRUE	https://eur-l	9	3	6	Directorate-General for Employment, Social
31991L0533	Council Direc	Consultation	0	["worker infc	1991	1	https://eur-l	L	FALSE	["Social provisions", "Approxim	TRUE	https://eur-l	7	2	5	Directorate-General for Employment, Social
31991L0368	Council Direc	Cooperation	0	["occupation	1991	2	https://eur-l	L	TRUE	["Internal market - Principles",	FALSE	https://eur-l	9	2	6	DG III ,Äì Industry
31991L0534	Council Direc	Council Only	0	["agricultura	1991	1	https://eur-l	L	TRUE	["Social provisions", "Informati	FALSE	https://eur-l	2	1	0	Legal service
31991L0382	Council Direc	Cooperation	0	["occupation	1991	2	https://eur-l	L	TRUE	["Safety at work and elsewhere	FALSE	https://eur-l	7	3	5	Directorate-General for Employment, Social
31992R2079	Council Regu	Cooperation	0	["early retire	1992	4	https://eur-l	R	TRUE	["Social provisions", "Agricultur	FALSE	https://eur-l	4	2	4	Directorate-General for Agriculture and Rura
31992L0057	Council Direc	Cooperation	0	["occupation	1992	2	https://eur-l	L	FALSE	["Internal market - Principles",	TRUE	https://eur-l	10	3	6	Directorate-General for Employment, Social
31992L0058	Council Direc	Cooperation	0	["occupation	1992	2	https://eur-l	L	TRUE	["Internal market - Principles",	TRUE	https://eur-l	10	3	4	Directorate-General for Employment, Social
31992R2434	Council Regu	Cooperation	0	["free mover	1992	1	https://eur-l	R	TRUE	["Free movement of workers"]	FALSE	https://eur-l	7	2	4	Directorate-General for Employment, Social
31992L0056	Council Direc	Consultation	0	["worker infc	1992	1	https://eur-l	L	TRUE	["Social provisions", "Approxim	FALSE	https://eur-l	6	2	4	Directorate-General for Employment, Social
31992R1247	Council Regu	Consultation	1	["disabled pe	1992	1	https://eur-l	R	TRUE	["Social security for migrant wo	FALSE	https://eur-l	4	2	1	Directorate-General for Employment, Social
31992R3654	Council Regu	Council Only	0	["discount sa	1992	4	https://eur-l	R	TRUE	["Milk products"]	FALSE	https://eur-l	2	1	0	Directorate-General for Agriculture and Rura
31992L0104	Council Direc	Cooperation	0	["occupation	1992	2	https://eur-l	L	FALSE	["Approximation of laws", "Saf	TRUE	https://eur-l	11	2	6	Directorate-General for Employment, Social
31992L0091	Council Direc	Cooperation	0	["humanisat	1992	2	https://eur-l	L	FALSE	["Safety at work and elsewhere	TRUE	https://eur-l	12	2	3	Directorate-General for Employment, Social
31992L0029	Council Direc	Cooperation	0	["occupation	1992	2	https://eur-l	L	FALSE	["Transport", "Safety at work a	TRUE	https://eur-l	10	3	7	Directorate-General for Employment, Social
31992L0085	Council Direc	Cooperation	0	["humanisat	1992	2	https://eur-l	L	FALSE	["Safety at work and elsewhere	TRUE	https://eur-l	11	3	8	Directorate-General for Employment, Social
31993L0104	Council Direc	Cooperation	0	["humanisat	1993	2	https://eur-l	L	FALSE	["Internal market - Principles",	FALSE	https://eur-l	9	3	5	Directorate-General for Employment, Social
31993R1947	Council Regu	Consultation	0	["board of di	1993	4	https://eur-l	R	TRUE	["Provisions under Article 235 E	FALSE	https://eur-l	5	1	3	Directorate-General for Budget
31993L0088	Council Direc	Cooperation	0	["occupation	1993	2	https://eur-l	L	TRUE	["public health", "Safety at wor	FALSE	https://eur-l	7	3	4	Directorate-General for Employment, Social
31993L0103	Council Direc	Cooperation	0	["in-service t	1993	2	https://eur-l	L	FALSE	["Approximation of laws", "Fish	TRUE	https://eur-l	9	3	4	Directorate-General for Employment, Social
31994L0045	Council Direc	Cooperation	0	["group of cc	1994	1	https://eur-l	L	FALSE	["Approximation of laws", "Soc	FALSE	https://eur-l	8	3	4	Directorate-General for Employment, Social
31994L0033	Council Direc	Cooperation	0	["labour star	1994	2	https://eur-l	L	FALSE	["Approximation of laws", "Soc	TRUE	https://eur-l	8	3	7	Directorate-General for Employment, Social
31994R2063	Council Regu	Consultation	0	["vocational	1994	1	https://eur-l	R	TRUE	["Provisions governing the Insti	FALSE	https://eur-l	3	1	2	Task Force for the Accession Negotiations
31994R3096	Council Regu	Council Only	0	["discount sa	1994	4	https://eur-l	R	TRUE	["European Agricultural Guidan	FALSE	https://eur-l	2	1	0	Directorate-General for Agriculture and Rura
31994L0058	Council Direc	Cooperation	0	["vocational	1994	2	https://eur-l	L	FALSE	["Internal market - Principles",	FALSE	https://eur-l	6	3	2	DG VII ,Äì Transport

Data Collection Principles

Data Transparency & Reproducibility

- Store the data in open formats (best: ".csv"), avoid software-specific formats
- Document everything!
  - Variable codebook
  - `R` scripts / do files
- Have a master dataset that you **never** overwrite
  - data you can always go back to
- backup your files! (Dropbox / Google Drive)

Data Collection Principles

Example Variable Codebook

General Questions on European Integration *CHES 2019 Codebook*

EU_POSITION = overall orientation of the party leadership towards European integration in 2019.

1 = Strongly opposed
2 = Opposed
3 = Somewhat opposed
4 = Neutral
5 = Somewhat in favor
6 = In favor
7 = Strongly in favor

EU_POSITION_SD = standard deviation of expert placement of overall orientation of the party leadership towards European integration in 2019.

EU_SALIENCE = relative salience of European integration in the party's public stance in 2019.

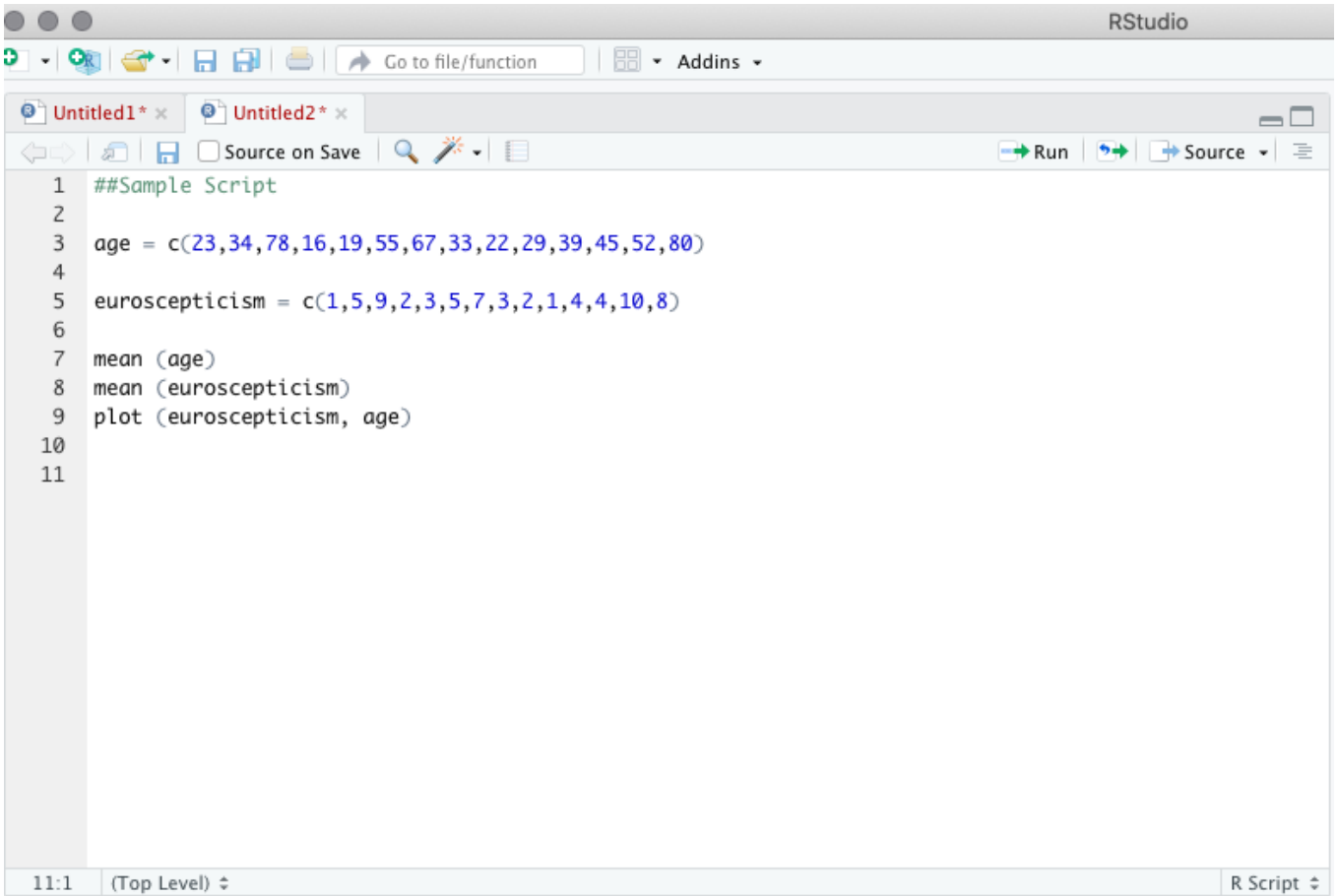
0 = European Integration is of no importance.
:
10 = European Integration is of great importance.

EU_DISSENT = degree of dissent on European integration in 2019.

0 = Party was completely united.
:
10 = Party was extremely divided.

Data Collection Principles

Example R Script



That's why using RMarkdown is great!

Ethical issues

- do not violate copyrights/terms of use: check if you have permission
- cite/give appropriate credit for data
- Respect privacy: anonymise!
- Do not coerce: make sure you have consent

Markup Languages

Web page Structures

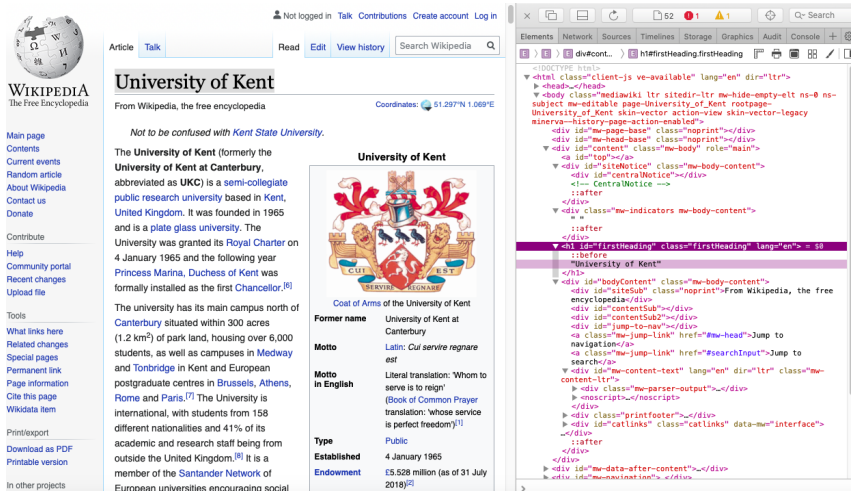
- **HTML** *Hypertext Markup Language*: series of symbols that define the structure and presentation of a web page. Invaluable to learn in web scraping as this is the standard language.
- **XML** *Extensible Markup Language*: almost identical to HTML but more flexible since tags can be user-defined.
- **JSON** *JavaScript Object Notation*: more lightweight formatting language, no start/end tags, just key/value pairs and curly or square brackets to express hierarchies.

HTML

Figure 2.1 Browser view of a simple HTML document

Figure 2.2 Source view of a simple HTML document

HTML: Element Inspector



HTML: Core Elements

- **head**: metadata element, usually containing document characteristics
- **title**: document title, used by search engines
- **body**: contains the webpage contents
- **header**: defines section headers
- **div**: defines a section of the html document
- **link**: defines links to external resources. **Attributes** likw rel="" and href="" define, respectively, type of link and url of the external resource

HTML: Formatting Tags

- **p**: defines separate paragraphs
- **br**: defines a line break
- **b**: bold text
- **i**: text in italic
- **ul**: unordered lists
- **ol**: ordered lists
- **h1 ... h6**: headline size
- **span**: usually combined with attribute class="" or style="" marks up parts of text to change colors and style
- **table**: used to begin a table **tr** used to begin a row, **td** for cells, and **th** for header cells.

Always preceded and followed starting and closing brackets. For more information see `https://www.w3schools.com/html/default.asp`.

XML & JSON: Examples

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<bond_movies>
  <movie id="1">
    <name>Dr. No</name>
    <year>1962</year>
    <actors bond="Sean Connery" villain="Joseph Wiseman"/>
    <budget>1.1M</budget>
    <boxoffice>59.5M</boxoffice>
  </movie>
  <movie id="2">
    <name>Live and Let Die</name>
    <year>1973</year>
    <actors bond="Roger Moore" villain="Yaphet Kotto"/>
    <budget>7M</budget>
    <boxoffice>126.4M</boxoffice>
  </movie>
  <movie id="3">
    <name>Skyfall</name>
    <year>2012</year>
    <actors bond="Daniel Craig" villain="Javier Bardem"/>
    <budget>175M</budget>
    <boxoffice>1108.6M</boxoffice>
  </movie>
</bond_movies>
```

Figure 3.1 An XML code example: James Bond movies

```json
{"indy movies" :[
      {
      "name" : "Raiders of the Lost Ark",
      "year" : 1981,
      "actors" : {
            "Indiana Jones": "Harrison Ford",
            "Dr. René Belloq": "Paul Freeman"
            },
      "producers": ["Frank Marshall", "George Lucas", "Howard Kazanjian"],
      "budget" : 18000000,
      "academy_award_ve": true
      },
      {
      "name" : "Indiana Jones and the Temple of Doom",
      "year" : 1984,
      "actors" : {
            "Indiana Jones": "Harrison Ford",
            "Mola Ram": "Amish Puri"
            },
      "producers": ["Robert Watts"],
      "budget" : 28170000,
      "academy_award_ve": true
      },
      {
      "name" : "Indiana Jones and the Last Crusade",
      "year" : 1989,
      "actors" : {
            "Indiana Jones": "Harrison Ford",
            "Walter Donovan": "Julian Glover"
            },
      "producers": ["Robert Watts", "George Lucas"],
      "budget" : 48000000,
      "academy_award_ve": false
      }]
}
```

Figure 3.9 JSON code example: Indiana Jones movies

Steps of Web Scraping

1. **Inspect** background structure of the web page
2. **Read** HTML/XML/JSON structure into R
3. **Parse** = extract information from relevant webpage elements only & convert to usable format (data frame)

Exercise 1: Inspecting Web-Pages

- Open a webpage (e.g. newspaper or Wikipedia).
- Have a look at the source code (right-click on the webpage with your mouse and select the command that mentions 'page source' or 'inspect element').
- Check out the structure of the code underlying the webpage: can you find some familiar element?

Exercise 2: Create a Web-Page

1. Write up a Word Document with a section head in bold and a series of paragraphs and an ordered list (you can also copy/paste content from a website)
2. Save the word document as an html file (scroll down the 'file format' field, after clicking 'save as').
3. Open the html file and explore it with the 'Inspect Element' function in your browser (right-clicking). What does the HTML structure of your document look like?

Exercise 3: Re-create a Web-Page Structure

1. Go to: w3schools.com/html/tryit
2. Play with HTML editor and try to re-create the main section of this webpage (doesn't have to be an identical/perfect result, try to approximate it as best you can!): e.g.EU Treaties

The w3schools website also has suggestions on useful tags you can use, with examples, see:
`https://www.w3schools.com/html/default.asp`)