Data Science 101 Lecture Series

October 22, 2025

ICTR Institute for Clinical and Translational Research AT EINSTEIN AND MONTEFIORE

Building Bridges in the Bronx and Beyond

DATA SCIENCE INSTITUTE AT EINSTEIN

What is Data Science?

*Data Science is an **interdisciplinary** field that combines statistics, computer science, and domain knowledge to extract insights and make predictions from data*

Core Data Science Elements:

- **Data Wrangling** → cleaning, organizing, transforming raw data
- **Statistics and Machine Learning** → methods to model uncertainty, find and test patterns in the data, and make predictions
- **Programming** (R, SAS, C, Python,…) → software used for analysis
- **Domain Expertise** → meaningful and actionable results

Why is Data Science important?

- **Data wrangling** - ensures analyses are built on clean, consistent inputs
- **Transparent methods** - coding in R/Python/SAS make results reproducible
- **Statistical rigor** - guards against false positives, p-value hacking, and overfitting
- **Generalizable models** - improve confidence that findings hold beyond a single dataset

Why is Data Science Important?

REPORT

Genetic Signatures of Exceptional Longevity in Humans

PAOLA SEBASTIANI, NADIA SOLOVIEFF, ANNIBALE PUCA, STEPHEN W. HARTLEY, EFTHYMIA MELISTA, STACY ANDERSEN, DANIEL A. DWORKIS, JEMMA B. WILK, RICHARD H. MYERS, [...], AND THOMAS T. PERLS +3 authors Authors Info & Affiliations

SCIENCE - 1 Jul 2010 - Vol 329, Issue 5987 - DOI: 10.1126/science.1190532

178 59

This article has been retracted.

Retraction
22 JUL 2011

Abstract

Healthy aging is thought to reflect the combined influence of environmental factors (lifestyle choices) and genetic factors. To explore the genetic contribution, we

CURRENT ISSUE

Why is Data Science Important?

Obesity
A Research Journal

THE OBESITY SOCIETY

THIS ARTICLE HAS BEEN RETRACTED | Free Access

Retraction statement: LA sprouts randomized controlled nutrition, cooking and gardening program reduces obesity and metabolic risk in Latino youth

First published: 02 November 2015 | https://doi.org/10.1002/oby.21390 | Citations: 6

PDF TOOLS SHARE

Abstract

The above article, published online on 9 May 2015 in Wiley Online Library (wileyonlinelibrary.com), and in Volume 23, pp. 1244-1251, has been retracted by agreement between the authors, the journal Editors-in-Chief, Eric Ravussin and Donna Ryan, the Obesity Society, and Wiley Periodicals, Inc. The retraction has been agreed to because the statistical analysis was not correct given the cluster-randomized design, and the wrong degrees of freedom were used. The conclusion that the original paper drew about having demonstrated treatment efficacy was not supported in the corrected

Obesity

Volume 23, Issue 12
December 2015
Page 2522

OBESITY SOCIETY
Research. Education. Action.

Related Information

Recommended

Retracted: LA sprouts randomized controlled nutrition and gardening program reduces obesity and metabolic risk in latino youth

Nicole M. Gatto, Lauren C. Martinez, Donna Spruijt-Metz, Jaimie N. Davis

Obesity

Why is Data Science Important?

Erratum to "Correlation not Causation: The Relationship between Personality Traits and Political Ideologies" [American Journal of Political Science 56, (1), 34-51]

Brad Verhulst, Lindon Eaves, Peter K. Hatemi

Political Science, Huck Institutes of the Life Sciences

Research output: Contribution to journal › Comment/debate › peer-review

Reverse-coded variables:
Researchers used incorrectly coded variables in their analysis causing their conclusions to be the complete opposite of what the data actually supported.

Good Data Science leads to Scientific Advancement

"The bottom line…..is that when **science is done well**, it produces believable, **replicable**, and **generalizable** findings"

-Jon Krosnick, the Frederic O. Glover Professor of Humanities and Social Sciences in the Stanford School of Humanities and Sciences

Why this series?

To bridge *theory* and *application*:

- Lunch & Learn sessions → *concepts + discussion*
- R workshops → *hands-on practice (Spring, 2026)*

Fall 2025

DS 101 lecture series

Date	Session
10/22 12-1:30	Lunch and Learn: Data wrangling
10/29 12-1:30	Lunch and Learn: Concepts in Biostatistics
11/5 12-1:30	Lunch and Learn: Machine Learning
11/12 12-1:30	Lunch and Learn: Electronic Health Records
11/19 12-1:30	Lunch and Learn: Omics data analysis

Why is Data Wrangling our first lecture?

- Admittedly, not the most exciting part of data science compared to Machine Learning or Omics
- But it is the absolute **foundation**:
  - Every analysis depends on clean, reliable data.
  - Without it, even the most complex and sophisticated models will fail.
- In this lecture you will:
  - Learn how to recognize and handle messy, real-world data.
  - Build skills that you'll apply in **every other session** (biostats, ML, EHR , omics).

Data Wrangling: Making Messy Data Usable

Juan Lin

Department of Epidemiology and Population Health

Oct. 22,2025

Author Manuscript

Author Manuscript

Author

HHS Public Access

Author manuscript

Am J Obstet Gynecol. Author manuscript; available in PMC 2021 March 01.

Published in final edited form as:

Am J Obstet Gynecol. 2020 March ; 222(3): 269.e1–269.e8. doi:10.1016/j.ajog.2019.10.005.

Pregnant? Validity of the Pregnancy Checkbox on Death Certificates in Four States, and Characteristics Associated with Pregnancy Checkbox Errors

Ms. Andrea CATALANO, MPH[1], Nicole L. DAVIS, PhD[1], Emily E. PETERSEN, MD[1], Mr. Christopher HARRISON, MPH[2], Lyn KIELTYKA, PhD[1,3], Ms. Mei YOU, MS[4], Elizabeth J. CONREY, PhD, RD[1,5], Mr. Alexander C. EWING, MPH[1], William M. CALLAGHAN, MD[1], David GOODMAN, PhD[1]

[1]Division of Reproductive Health, Centers for Disease Control and Prevention

[2]Office of Vital Records, Georgia Department of Public Health

[3]Bureau of Family Health, Louisiana Office of Public Health

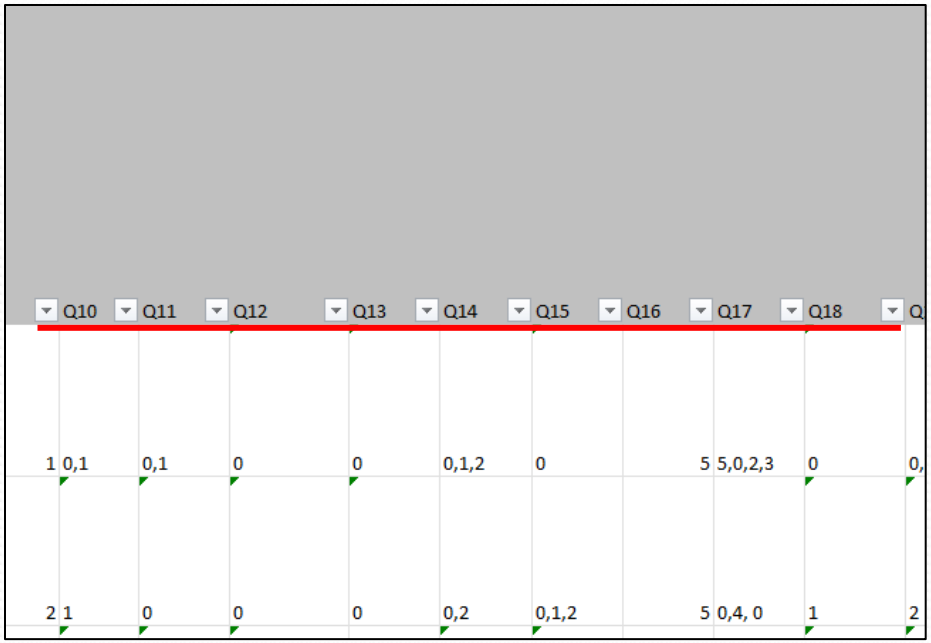
[4]Division of Vital Records and Health Statistics, Michigan Department of Health and Human Services

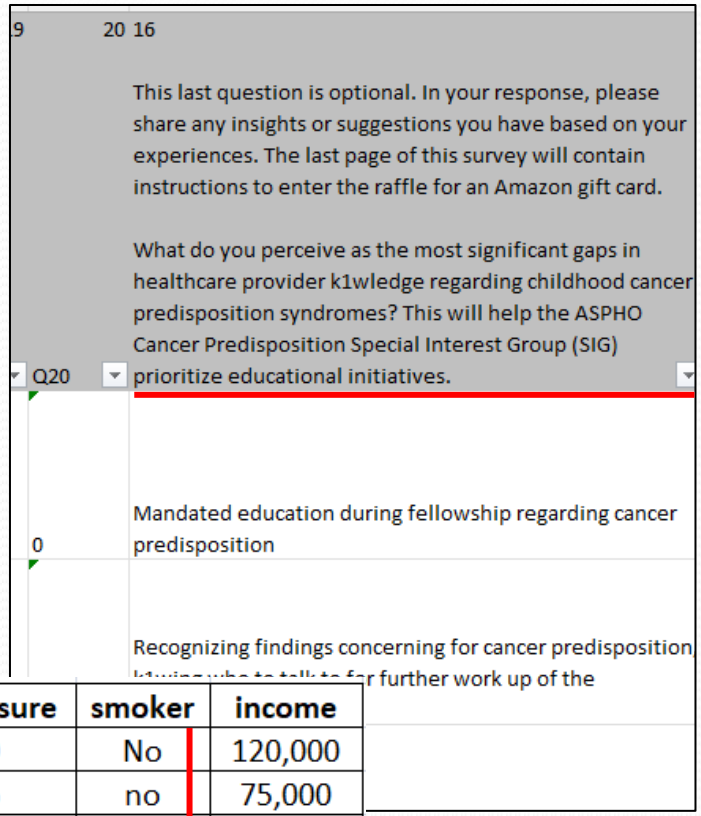
[5]Bureau of Maternal, Child and Family Health, Ohio Department of Health

Abstract

Background: Maternal mortality rates in the United States appear to be increasing. One potential reason may be increased identification of maternal deaths after the addition of a pregnancy checkbox to the death certificate. In 2016, four state health departments (Georgia, Louisiana,

Survey Data





respondent_id	age	occupation	visit date	blood pressure	smoker	income
1	34	Physician	3/15/2023	120/80	No	120,000
2	29	nurse	15/03/2023	118/76	no	75,000
3	41	Medical Assistant	3/20/2023	140/90	yes	65,000
4	33		23-Mar-23	125/85	n/a	90,000
5	52	Declined	3/25/2023	135/89	No	$125,000
6	33	Nurse (Part-Time)	3/28/2023	122\85	No	N/A

Biomedical data

A

Date

12/23/21

3/17/22

4/6

4/12

S

DIF Performed? 0 = No, 1 = Yes

0

1

0

1 - There is no deposition of IgG, IgM, IgA, C3, or fibrinogen detected in the epidermis, dermis, or basement

frontiers | Frontiers in Bioinformatics

TYPE Hypothesis and Theory
PUBLISHED 05 January 2024
DOI 10.3389/fbinf.2023.1328613

Check for updates

OPEN ACCESS

EDITED BY
Wei Jiang,
First Affiliated Hospital of Fujian Medical University, China

REVIEWED BY
Yi Han,
University of Texas Southwestern Medical Center, United States
Mingyao Pan,
University of Pennsylvania, United States
Yoshihiro Noguchi,
Gifu Pharmaceutical University, Japan

*CORRESPONDENCE
Wen Zou,
wen.zou@fda.hhs.gov

RECEIVED 27 October 2023
ACCEPTED 11 December 2023
PUBLISHED 05 January 2024

CITATION
Le H, Chen R, Harris S, Fang H, Lyn-Cook B, Hong H, Ge W, Rogers P, Tong W and Zou W (2024), RxNorm for drug name normalization: a case study of

RxNorm for drug name normalization: a case study of prescription opioids in the FDA adverse events reporting system

Huyen Le[1], Ru Chen[2], Stephen Harris[1], Hong Fang[3], Beverly Lyn-Cook[4], Huixiao Hong[1], Weigong Ge[1], Paul Rogers[1], Weida Tong[1] and Wen Zou[1]*

[1]Division of Bioinformatics and Biostatistics, Jefferson, AR, United States, [2]Office of Translational Science, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, United States, [3]Office of Scientific Coordination, Jefferson, AR, United States, [4]Division of Biochemistry Toxicity, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, United States

Numerous studies have been conducted on the US Food and Drug Administration (FDA) Adverse Events Reporting System (FAERS) database to assess post-marketing reporting rates for drug safety review and risk assessment. However,

Common Problems in Raw Data

- Missing values
- Inconsistent formats
- Categorical chaos
- Typos and spelling variants
- Duplicates
- Mismatched IDs in linked datasets

Missing Data: Types

- The types of missingness
  - Partial missingness
  - Structural missingness (due to logic skips)
  - Dropout in longitudinal study
- Representations
  - Blanks
  - Placeholders (*"NA", "N/A", "n.a.","999"*)
  - Hidden text(*"refused to answer", "not applicable", "unknown")*

Missing Data: Fixes

- Fix
  - Identify missingness
  - Standardize missing codes
  - Assess if missingness is informative
- Decide what to do next
  - Exclude
  - Impute
  - Flag for stratified analysis

Inconsistent formats

- What does it mean?
  - Dates : *“03/10/25", "10-Mar-2025", "20250310", “10/03/25”*
  - Numerical and text variables: *"50“vs"$50“, "1000" vs "1,000“,“3” vs “>2”*
  - Identifiers: *"00123" vs "123"; confusing the letter “O” with the number “0”*
  - Units / Measurements: *“180 mg/dL” vs 180; “5.6 mmol/L” vs “101 mg/dL” (in same column)*

Fix inconsistent formats

- Format Column names
  - Recommended format: lowercase letters; avoid special characters; short but meaningful
    - Examples: Age at diagnosis → age_dx; DOB (Date of Birth) → dob; Cholesterol(mg/dL) →chol
- Format Columns
  - Sort and filter to reveal inconsistent formats
  - Pick one format
  - Standardize (remove hidden spaces, currency, commas, or percent symbols)

Categorical chaos and typos

- Examples
  - Same categories coded inconsistently: *"Male"*, *"male"*, *"M"*, *"m"*
  - Typos: “diabetes”, “daibetes”, “diabtes”, "diabetees"
  - Inconsistent casing: “COVID-19”, “covid19”, “COVID19”
- Fix
  - Group and review unique values
  - Re-check categories and frequencies after recoding to ensure consistency
  - Validate data ranges and flag outliers

Duplicates and mismatched IDs

- **Retracted article**
  See the retraction notice

  > Nat Med. 2006 Nov;12(11):1294-300. doi: 10.1038/nm1491. Epub 2006 Oct 22.

  ## Genomic signatures to guide the use of chemotherapeutics

  Anil Potti [1], Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster, Joseph R Nevins

  Affiliations + expand

  PMID: 17057710 DOI: 10.1038/nm1491

  Erratum in

  Nat Med. 2007 Nov;13(11):1388

  Nat Med. 2008 Aug;14(8):889

- *, visit*
- *in*
- *in one*
- tes and

Excel and R for data wrangling

Feature	Excel	R
Ease of Use	Intuitive GUI	Steeper learning curve
Speed with Large Data	Slower, limited by memory	Fast and efficient
Reproducibility	Hard to track manual steps	Fully scriptable and reproducible
Error-prone?	High (manual edits, copy/paste)	Lower (if code is correct)
Visualization	Charts, conditional formatting	ggplot2, custom plots
Good For...	Quick reviews, small datasets	Automation, large/complex data

This session

Next session: Introduction to Tidyverse and data wrangling in R

Case Study: Survey Data

Respondent_ID	Q1	Q2	Time-stamp
R001	1	Agree	2023-07-01 08:30
R007	1,3	Agree	2023-07-01 10:00
R002	2	Strongly Agree	2023-07-01 08:45
R003	refuse to answer	Neutral	2023-07-01 09:00
R-004	1	Disagree	2023-07-01 09:15
r005	1,2,3	N/A	2023-07-01 09:30
R006	3	see above	2023-07-01 09:45
R010	NA	Disagree	2023-07-01 10:45
R008	2		2023-07-01 10:15
R009	3	Neutral	2023-07-01 10:30

I. survey answers

II. Participants Information

respondent_id	Full Name	Date of Birth	Gender	Group Code
R001	Alice Smith	01/02/1990	Male	A_1
R002	Bob Johnson	1991-03-04	female	A_1
R003	Charlie Lee	March 5, 1992	M	A1
R004	Dana White		F	B-2
R005	Eve Black	03-06-1993	FEMALE	B2
R006	Frank Wright	07/07/1990	m	B-2
R007	Grace Kim	N/A	Male	C3
R008	Hank Brown	1990/08/08	Other	C_3
R009	Ivy Zhao	09-09-1991	F	C-3
R010	John Wu	unknown	f	C3

I.

Respondent_ID	Q1	Q2	Time-stamp
R001	1	Agree	2023-07-01 08:30
R007	1,3	Agree	2023-07-01 10:00
R002	2	Strongly Agree	2023-07-01 08:45
R003	refuse to answer	Neutral	2023-07-01 09:00
R-004	1	Disagree	2023-07-01 09:15
r005	1,2,3	N/A	2023-07-01 09:30
R006	3	see above	2023-07-01 09:45
R010	NA	Disagree	2023-07-01 10:45
R008	2		2023-07-01 10:15
R009	3	Neutral	2023-07-01 10:30

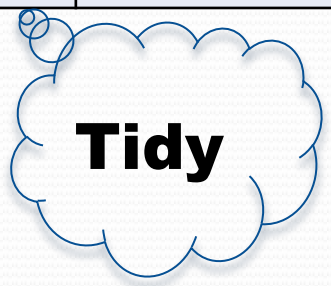
II.

respondent_id	Full Name	Date of Birth	Gender	Group Code
R001	Alice Smith	01/02/1990	Male	A_1
R002	Bob Johnson	1991-03-04	female	A_1
R003	Charlie Lee	March 5, 1992	M	A1
R004	Dana White		F	B-2
R005	Eve Black	03-06-1993	FEMALE	B2
R006	Frank Wright	07/07/1990	m	B-2
R007	Grace Kim	N/A	Male	C3
R008	Hank Brown	1990/08/08	Other	C_3
R009	Ivy Zhao	09-09-1991	F	C-3
R010	John Wu	unknown	f	C3

```
=IF(OR(C15="NA",C15="N/A",C15="see above",C15="unknown",C15="",C15="refuse to answer"),"",C15)
```

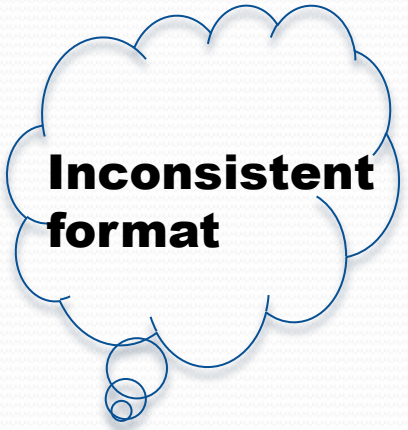
Q1	Q2	Date of Birth
1	Agree	01/02/1990
1,3	Agree	1991-03-04
2	Strongly Agree	March 5, 1992
refuse to answer	Neutral	
1	Disagree	03-06-1993
1,2,3	N/A	07/07/1990
3	see above	N/A
NA	Disagree	1990/08/08
2		09-09-1991
3	Neutral	unknown

Q1	Q2	Date of Birth
1	Agree	01/02/1990
1,3	Agree	1991-03-04
2	Strongly Agree	March 5, 1992
	Neutral	
1	Disagree	03-06-1993
1,2,3		07/07/1990
3		
	Disagree	1990/08/08
2		09-09-1991
3	Neutral	



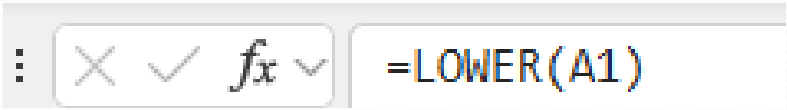
I.

Respondent_ID	Q1	Q2	Time-stamp
R001	1	Agree	2023-07-01 08:30
R007	1,3	Agree	2023-07-01 10:00
R002	2	Strongly Agree	2023-07-01 08:45
R003		Neutral	2023-07-01 09:00
R-004	1	Disagree	2023-07-01 09:15
r005	1,2,3		2023-07-01 09:30
R006	3		2023-07-01 09:45
R010		Disagree	2023-07-01 10:45
R008	2		2023-07-01 10:15
R009	3	Neutral	2023-07-01 10:30



II.

respondent_id	Full Name	Date of Birth	Gender	Group Code
R001	Alice Smith	01/02/1990	Male	A_1
R002	Bob Johnson	1991-03-04	female	A_1
R003	Charlie Lee	March 5, 1992	M	A1
R004	Dana White		F	B-2
R005	Eve Black	03-06-1993	FEMALE	B2
R006	Frank Wright	07/07/1990	m	B-2
R007	Grace Kim		Male	C3
R008	Hank Brown	1990/08/08	Other	C_3
R009	Ivy Zhao	09-09-1991	F	C-3
R010	John Wu		f	C3



Respondent_ID	Q1	Q2	Time-stamp

Tidy

respondent_id	q1	q2	timestamp

Tidy

respondent_id	Full Name	Date of Birth	Gender	Group Code

respondent_id	full_name	dob	gender	group_code

=IFERROR(DATEVALUE(H41),"")

Date of Birth
01/02/1990
1991-03-04
March 5, 1992
03-06-1993
07/07/1990
1990/08/08
09-09-1991

dob
1/2/1990
3/4/1991
3/5/1992
3/6/1993
7/7/1990
8/8/1990
9/9/1991

Tidy

Q1
1
1,3
2
1
1,2,3
3
2
3

Tidy

q1_1	q1_2	q1_3
1	0	0
1	0	1
0	1	0
1	0	0
1	1	1
0	0	1
0	1	0
0	0	1

Categorical chaos

II.

respondent_id	full_name	dob	gender	group_code
R001	Alice Smith	1/2/1990	Male	A_1
R002	Bob Johnson	3/4/1991	female	A_1
R003	Charlie Lee	3/5/1992	M	A1
R004	Dana White		F	B-2
R005	Eve Black	3/6/1993	FEMALE	B2
R006	Frank Wright	7/7/1990	m	B-2
R007	Grace Kim		Male	C3
R008	Hank Brown	8/8/1990	Other	C_3
R009	Ivy Zhao	9/9/1991	F	C-3
R010	John Wu		f	C3

```
=LOWER(IF(LOWER(R[-13]C)="m","male",IF(LOWER(R[-13]C)="f","female",R[-13]C)))
```

gender
Male
female
M
F
FEMALE
m
Male
Other
F
f

gender
male
female
male
female
female
male
male
other
female
female

Tidy

```
=TRIM(SUBSTITUTE(SUBSTITUTE(J67,"_",""),"-",""))
```

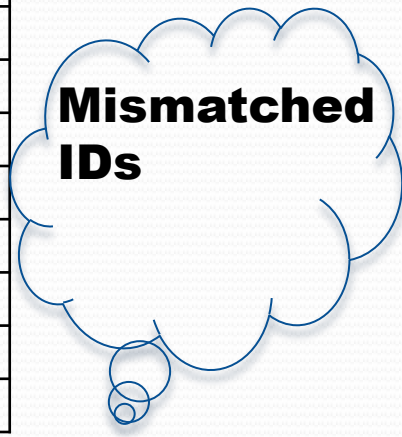
group_code
A_1
A_1
A1
B-2
B2
B-2
C3
C_3
C-3
C3

group_code
A1
A1
A1
B2
B2
B2
C3
C3
C3
C3

Tidy

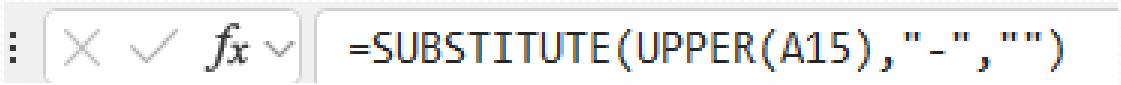
I.

respondent_id	q1_1	q1_2	q1_3	q2	timestamp
R001	1	0	0	Agree	2023-07-01 08:30
R007	0	1	0	Agree	2023-07-01 10:00
R002				Strongly Agree	2023-07-01 08:45
R003	1	0	0	Neutral	2023-07-01 09:00
R-004	1	1	1	Disagree	2023-07-01 09:15
r005	0	0	1		2023-07-01 09:30
R006	1	0	1		2023-07-01 09:45
R010	0	1	0	Disagree	2023-07-01 10:45
R008	0	0	1		2023-07-01 10:15
R009				Neutral	2023-07-01 10:30



II.

respondent_id	full_name	date_of_birth	gender	group_code
R001	Alice Smith	1/2/1990	male	A1
R002	Bob Johnson	3/4/1991	female	A1
R003	Charlie Lee	3/5/1992	male	A1
R004	Dana White		female	B2
R005	Eve Black	3/6/1993	female	B2
R006	Frank Wright	7/7/1990	male	B2
R007	Grace Kim		male	C3
R008	Hank Brown	8/8/1990	other	C3
R009	Ivy Zhao	9/9/1991	female	C3
R010	John Wu		female	C3

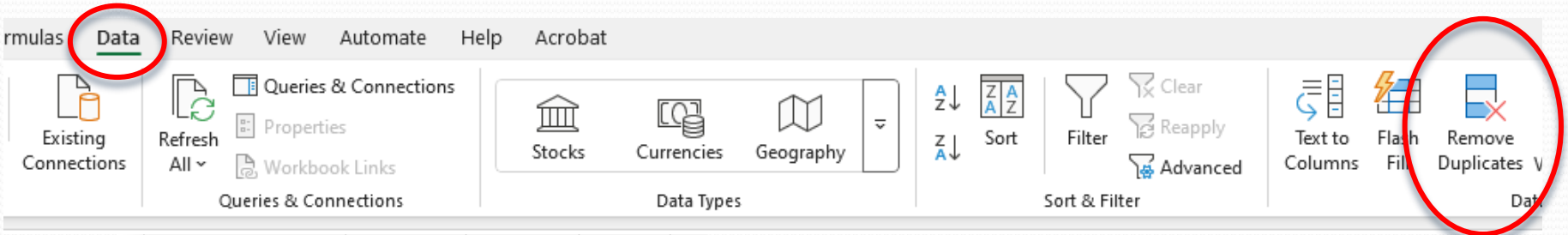


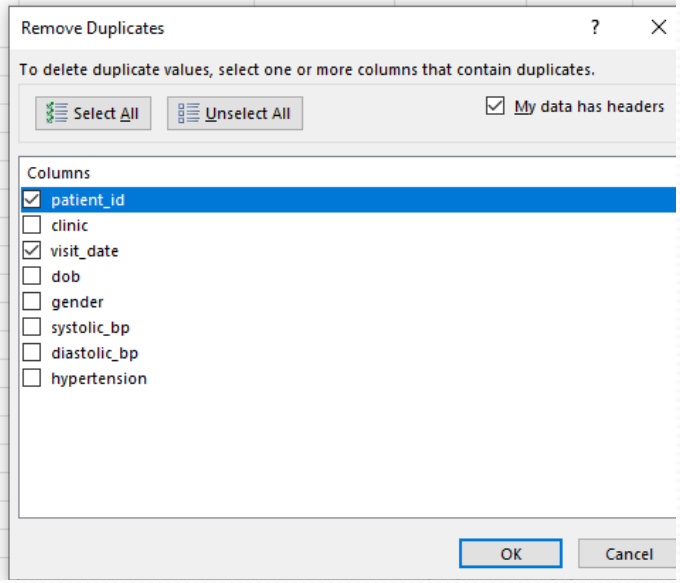
respondent_id
R001
R007
R002
R003
R-004
r005
R006
R010
R008
R009

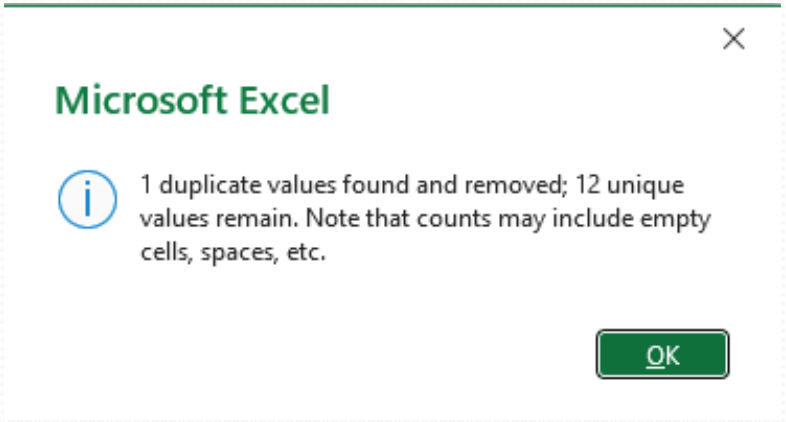
respondent_id
R001
R007
R002
R003
R004
R005
R006
R010
R008
R009

Tidy

patient_id	clinic	visit_date	dob	gender	systolic_bp	diastolic_bp	hypertension
A001	A	7/1/2023	5/12/1990	male	120	80	no
A002	A	7/2/2023	12/6/1985	male	125	82	no
A002	A	7/2/2023	12/6/1985	male	NA	NA	NA



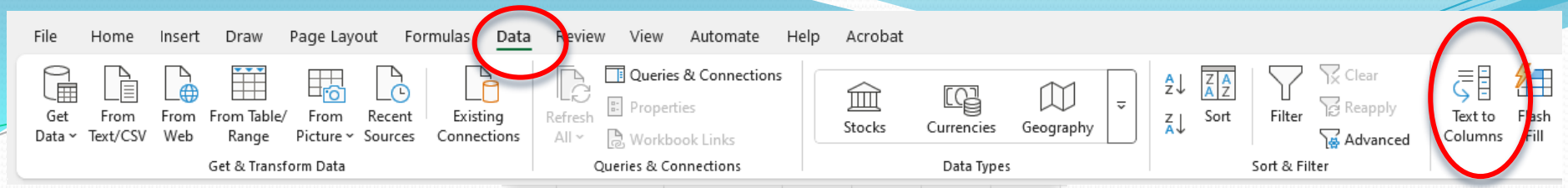


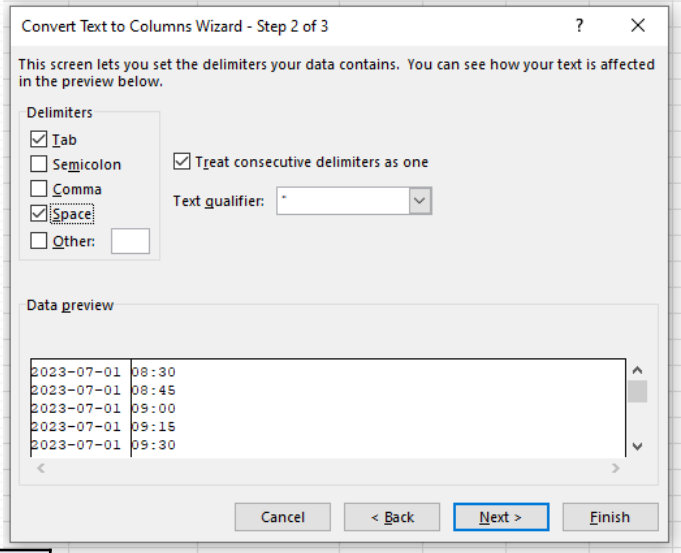


patient_id	clinic	visit_date	dob	gender	systolic_bp	diastolic_bp	hypertension
A001	A	7/1/2023	5/12/1990	male	120	80	no
A002	A	7/2/2023	12/6/1985	male	125	82	no

Structuring Data for Analysis

- Even once the data is clean, structure matters. Analysis often requires reshaping
  - Wide to long, long to wide,
  - Splitting columns (*'Full_name'* → *first_name, last_name*)
  - Merging datasets with consistent IDs





timestamp	full_name
2023-07-01 08:30	Alice Smith
2023-07-01 08:45	Bob Johnson
2023-07-01 09:00	Charlie Lee
2023-07-01 09:15	Dana White
2023-07-01 09:30	Eve Black
2023-07-01 09:45	Frank Wright
2023-07-01 10:00	Grace Kim
2023-07-01 10:15	Hank Brown
2023-07-01 10:30	Ivy Zhao
2023-07-01 10:45	John Wu

date	time	first_name	last_name
7/1/2023	8:30	Alice	Smith
7/1/2023	8:45	Bob	Johnson
7/1/2023	9:00	Charlie	Lee
7/1/2023	9:15	Dana	White
7/1/2023	9:30	Eve	Black
7/1/2023	9:45	Frank	Wright
7/1/2023	10:00	Grace	Kim
7/1/2023	10:15	Hank	Brown
7/1/2023	10:30	Ivy	Zhao
7/1/2023	10:45	John	Wu

I.

respondent_id	q1_1	q1_2	q1_3	q2	date	time
R001	1	0	0	Agree	7/1/2023	8:30
R007	0	1	0	Agree	7/1/2023	8:45
R002				Strongly Agree	7/1/2023	9:00
R003	1	0	0	Neutral	7/1/2023	9:15
R004	1	1	1	Disagree	7/1/2023	9:30
R005	0	0	1		7/1/2023	9:45
R006	1	0	1		7/1/2023	10:00
R010	0	1	0	Disagree	7/1/2023	10:15
R008	0	0	1		7/1/2023	10:30
R009				Neutral	7/1/2023	10:45

II.

respondent_id	first_name	last_name	dob	gender	group_code
R001	Alice	Smith	1/2/1990	male	A1
R002	Bob	Johnson	3/4/1991	female	A1
R003	Charlie	Lee	3/5/1992	male	A1
R004	Dana	White		female	B2
R005	Eve	Black	3/6/1993	female	B2
R006	Frank	Wright	7/7/1990	male	B2
R007	Grace	Kim		male	C3
R008	Hank	Brown	8/8/1990	other	C3
R009	Ivy	Zhao	9/9/1991	female	C3
R010	John	Wu		female	C3

respondent_id	q1_1	q1_2	q1_3	q2	date	time	first_name	last_name	dob	gender	group_code
R001	1	0	0	Agree	7/1/2023	8:30	Alice	Smith	1/2/1990	male	A1
R002	0	1	0	Strongly Agree	7/1/2023	8:45	Bob	Johnson	3/4/1991	female	A1
R003				Neutral	7/1/2023	9:00	Charlie	Lee	3/5/1992	male	A1
R004	1	0	0	Disagree	7/1/2023	9:15	Dana	White		female	B2
R005	1	1	1		7/1/2023	9:30	Eve	Black	3/6/1993	female	B2
R006	0	0	1		7/1/2023	9:45	Frank	Wright	7/7/1990	male	B2
R007	1	0	1	Agree	7/1/2023	10:00	Grace	Kim		male	C3
R008	0	1	0		7/1/2023	10:15	Hank	Brown	8/8/1990	other	C3
R009	0	0	1	Neutral	7/1/2023	10:30	Ivy	Zhao	9/9/1991	female	C3
R010				Disagree	7/1/2023	10:45	John	Wu		female	C3

Ready for analysis!

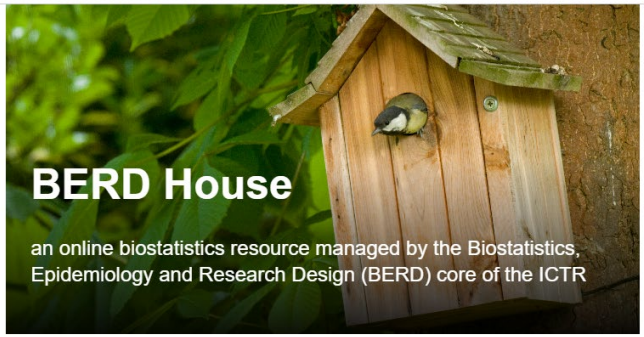
respondent_id	q1_1	q1_2	q1_3	q2	date	time	first_name	last_name	dob	gender	group_code
R001	1	0	0	Agree	7/1/2023	8:30	Alice	Smith	1/2/1990	male	A1
R002	0	1	0	Strongly Agree	7/1/2023	8:45	Bob	Johnson	3/4/1991	female	A1
R003				Neutral	7/1/2023	9:00	Charlie	Lee	3/5/1992	male	A1
R004	1	0	0	Disagree	7/1/2023	9:15	Dana	White		female	B2
R005	1	1	1		7/1/2023	9:30	Eve	Black	3/6/1993	female	B2
R006	0	0	1		7/1/2023	9:45	Frank	Wright	7/7/1990	male	B2
R007	1	0	1	Agree	7/1/2023	10:00	Grace	Kim		male	C3
R008	0	1	0		7/1/2023	10:15	Hank	Brown	8/8/1990	other	C3
R009	0	0	1	Neutral	7/1/2023	10:30	Ivy	Zhao	9/9/1991	female	C3
R010				Disagree	7/1/2023	10:45	John	Wu		female	C3

Best Practices

- Some quick survival tips
  - **Know your data's context** - Talk to clinicians and data owners — numbers alone can lie.
  - **Document everything** - You will not remember why you recoded '999' as 'NA' two months from now
  - **Keep originals** - Never overwrite raw data. You'll need it.
  - **Visual checks help** - Use plots, summaries, or tabulations to catch errors.
  - **Cross-check data with your team** - More eyes = fewer mistakes.

Biostatistics resources @ Einstein

- Einstein BERD house, online statistics resource (https://einsteinmed.edu/centers/ictr/biostatistics-epidemiology-research-design-core/berd-house/)



Melissa J. Fazzari, Ph.D., M.S.
BERD Director

ICTR Biostatistics Core

Biostatistics virtual walk-in clinic:

Meet with a biostatistician without an appointment and obtain quick advice on your project.

Every Tuesday virtual walk-in support between 3 - 5 pm via the following zoom link: https://einsteinmed.zoom.us/j/96403655408

BEST PRACTICES WHEN CREATING A DATA SET FOR ANALYSIS

1. **CREATING GOOD, DESCRIPTIVE COLUMN HEADERS using a SINGLE STRING (no spaces)**

WRONG WAY	BETTER WAY	COMMENT
# PILLS TAKEN/DAY	NUM_PILLS_PER_DAY	No special characters; Use underscores instead of spaces
CHEMO (Yes/No)	CHEMO	No Special characters; do not use parentheses

WHY THIS IS IMPORTANT: When the biostatistician reads your EXCEL data into SAS or another package, we rely on these column names to analyze the data correctly and efficiently. Sloppy names are confusing! Variables names should be long enough to be meaningful, but short enough to be easy to handle

Statistics, Machine Learning, and Data Science Training:

Getting Started

OMICS Data Analysis

Biostatistics Support and Guidelines

Who can I meet with to discuss my project?

The Division of Biostatistics offers biostatistics consulting and collaboration to enhance the quality and rigor of scientific research conducted by investigators at Einstein and Montefiore.

Longer-term collaborations:

Appointments with biostatistics faculty are available by appointment only.

How do I get my data set ready for analysis?

Study Design and Statistical Power

Journal Club

Announcements and Upcoming workshops

Wrap-Up

- Wrangling may not be glamorous, but it's the foundation of reliable analysis
- Most of your analysis time will be here
- Clean, well-structured data leads to solid science and fewer embarrassing mistakes

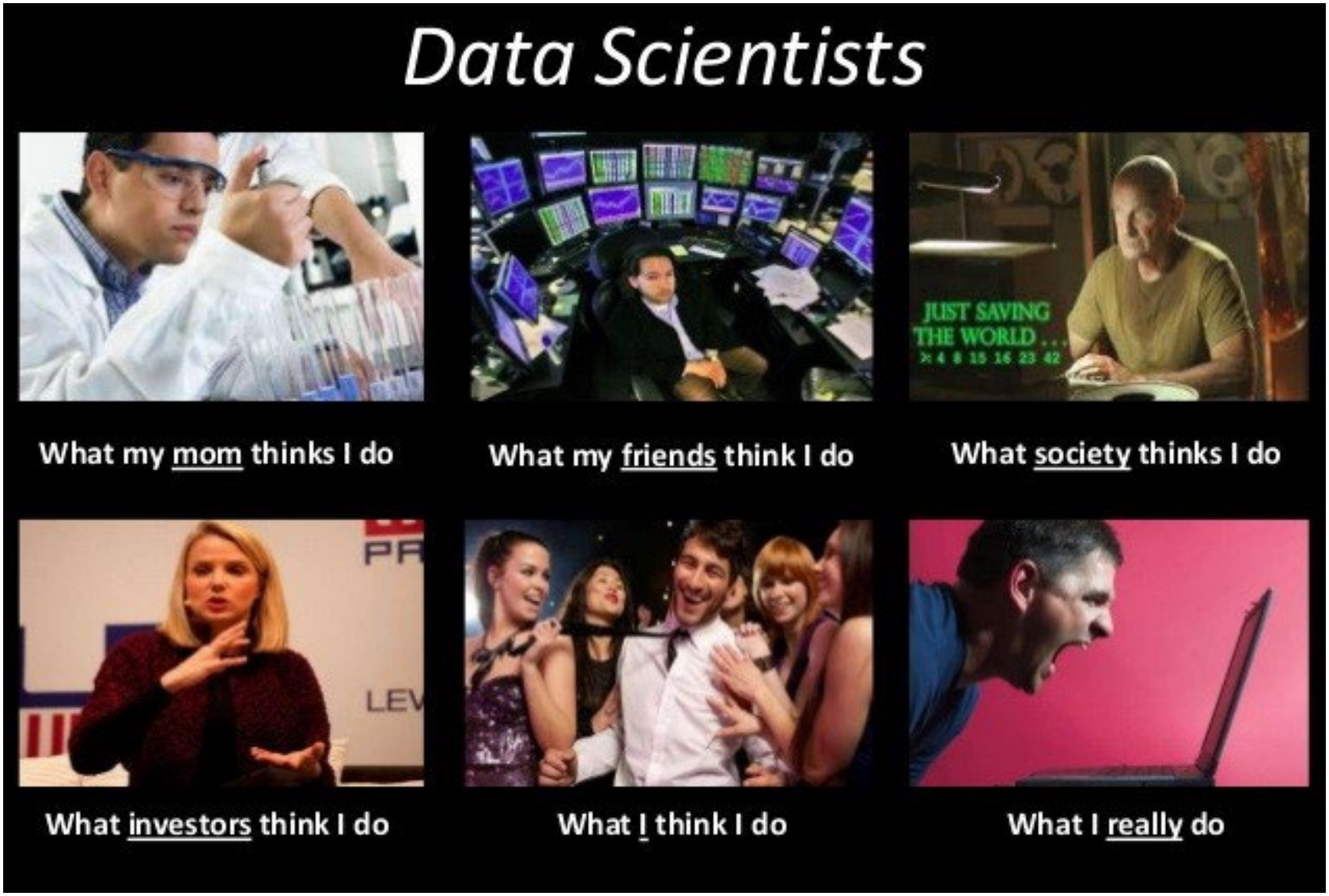


Image source: "Data Scientist – What People Think I Do / What I Really Do" meme, adapted from the original 'What People Think I Do' meme series (2012), circulated widely online.