Lesson 4: Data Types and Levels of Measurement

LSC 563: Data Visualization – Spring 2022

Class will Start at 5:15 (Zoom and in-person)

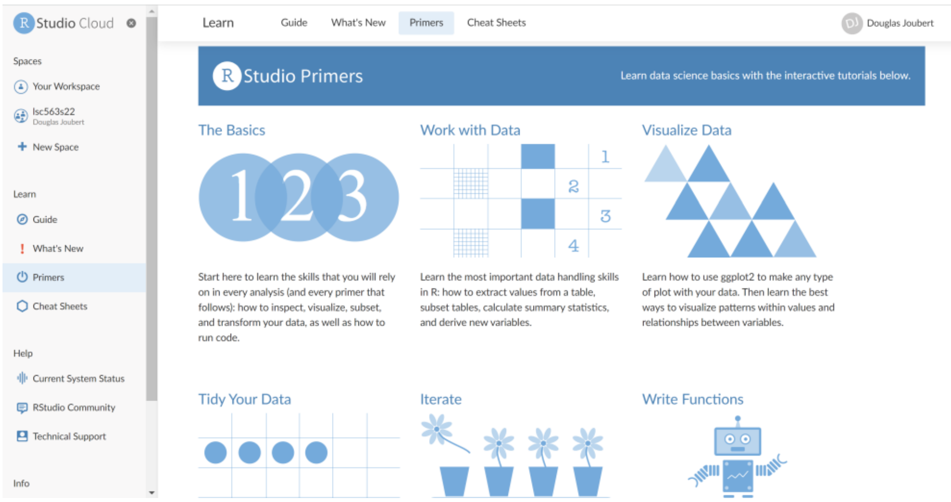
1

Check-In

- Blackboard – What is new and any issues
  - Schedule updated
  - All PDFs have been added under course documents (zipped folders)
  - New content added to lab folder
  - New content added to course content
- Anything else?

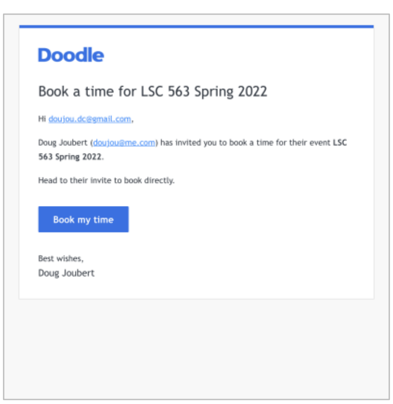
2

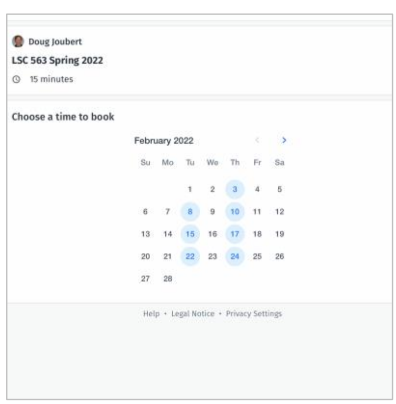
RStudio Cloud (Primers)

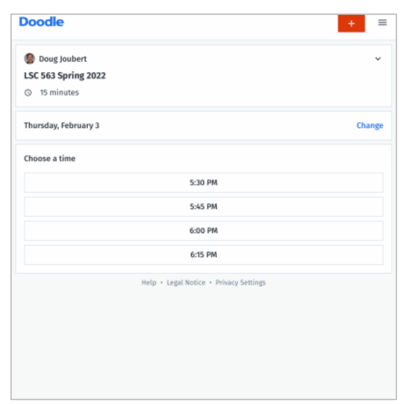


3

Virtual Office Hours

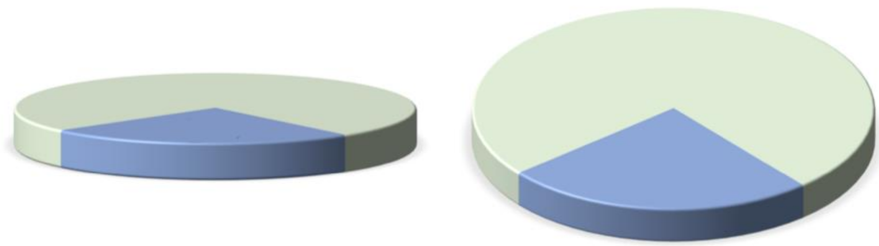






4

No 3-D: Example 1

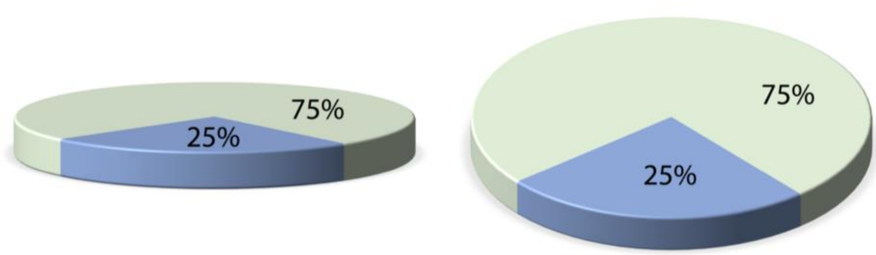


What are the % for each color?

Wilke, 2019

5

No 3-D: Example 2

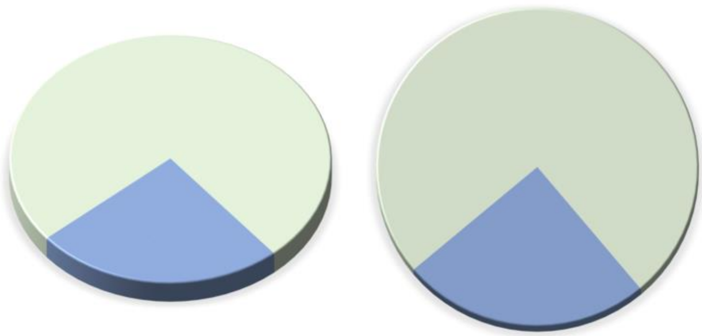


Even with the values, comparisons are hard

Wilke, 2019

6

No 3-D: Example 3

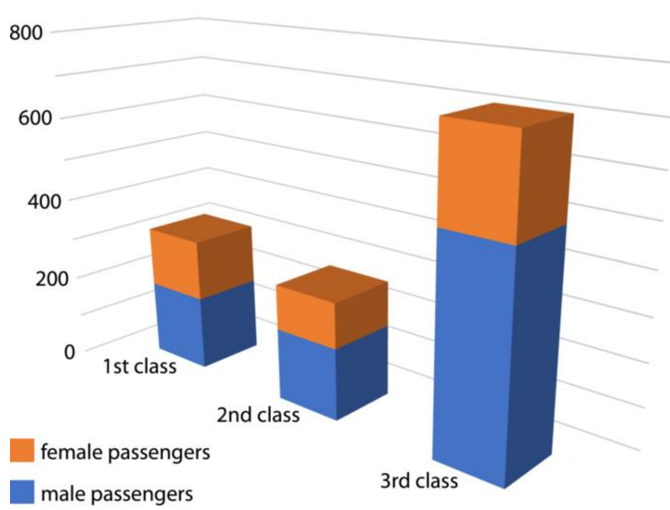


Better, but why would you want this option?

Wilke, 2019

7

No 3-D: Example 4



Wilke, 2019

8

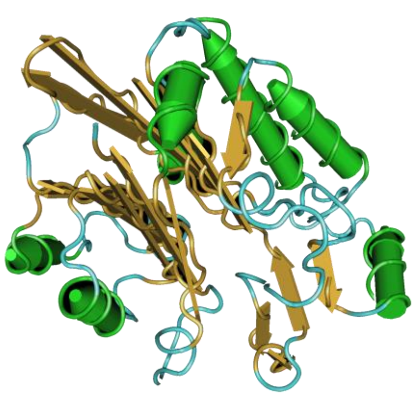
Appropriate use of 3D visualizations 1

What is different about this image?

ESRI, 2021

9

Appropriate use of 3D visualizations 2



1AKO: Exonuclease III from Escherichia Coli

Image source

10

Learning Objectives

- At the end of this lecture, students should be able to:
  - Distinguish between the three major datatypes
  - Describe and provide an example of data measured on a nominal, ordinal scale and interval scales
  - List the six types of questions that you might ask when working with a new dataset
  - List the recommendations for tidy data
  - Identify wide and long datasets
  - Compare and contrast aggregation and granularity

11

Datasets and Data Types

12

The Role of Data in Data Visualization

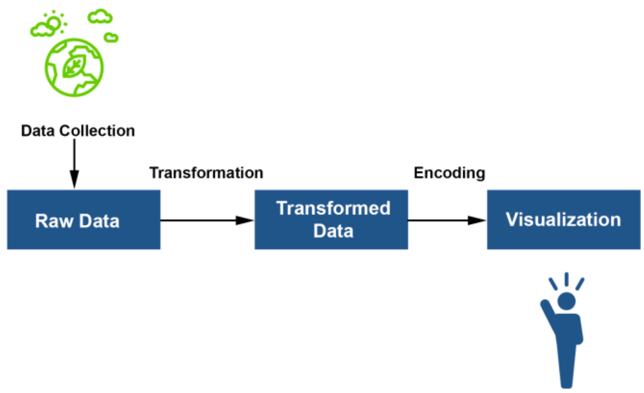
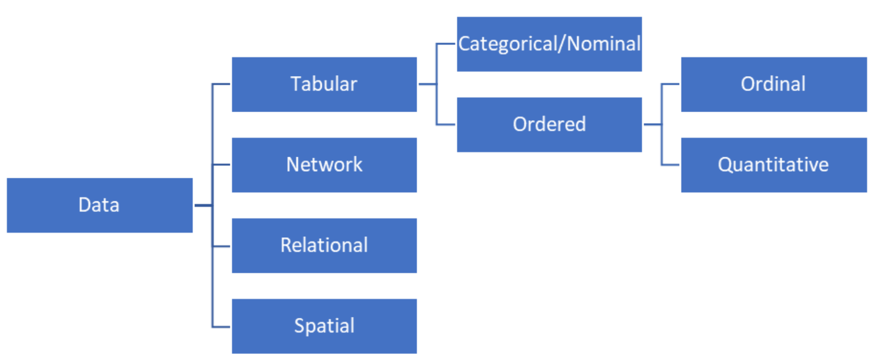


Image adapted from Bertini, E. (2016).

13

How to Think About Data



14

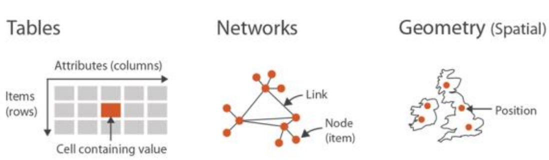
Role of Datasets and Data Types

- Goal of visualization is to transform data into a perceptually efficient graph
- Important to know about the different types of datasets and data
- First discuss dataset types and then we will discuss the various ways to describe or classify data.

Ware, 2013

15

Dataset Types



- Any collection of information that is the target of analysis
- Three basic dataset types that we will focus on:
  - Tables
  - Network
  - Spatial geometry

Munzner, 2015

16

Dataset Types: Tables

- Made up of rows and columns.
- Simple flat table
  - Each row represents an item of data
  - Each column is an attribute of the dataset.
- Each cell in the table is the combination of a row and a column and contains a value for that pair (R,C).

Munzner, 2015

17

Dataset Types: Tables - Example

Column (attribute)

OrderID	OrderDate	OrderPriority	ProductContainer	ProductBaseMargin	ShipDate
3	10/142006	5-low	Large	.08	10/21/2006
6	2/21/2018	4-Not Specified	Small	0.55	2/25/2018
32	6/5/2017	2-high	Medium	1.28	6/5/2017

Row (item)

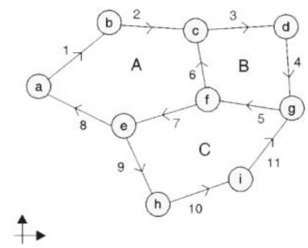
Cell (R x C)

Munzner, 2015

18

Dataset Types: Networks

- For specifying a relationship between two or more items
- Common plotting technique is the 'node-link' diagrams:
  - Nodes are items that make up a network, such as individuals in a social network
  - Edges are links that connect each node. The links describe the presence or absence of connections among the nodes



Directed Arcs, Nodes, and Areas

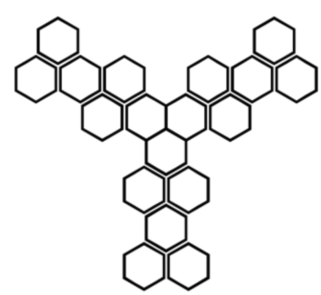
Arc	Start Node	End Node	Left Area	Right Area
1	a	b	X	A
2	b	c	X	A
3	c	d	X	B
4	d	g	X	B
5	g	f	C	B
6	f	c	A	B
7	f	e	C	A
8	e	a	X	A
9	e	h	C	X
10	h	i	C	X
11	i	g	C	X

Munzner, 2015

19

Dataset Types: Geometry

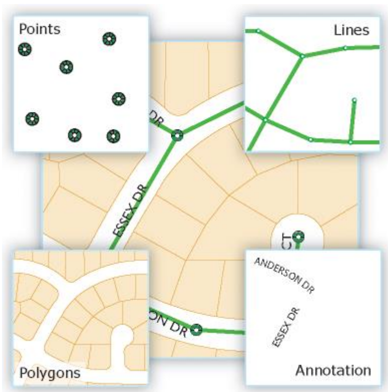
- Specifies information about the shape of items with explicit spatial positions:
  - Points
  - One-dimensional lines or curves,
  - 2D surfaces or regions, or 3D volumes



Munzner, 2015

20

Dataset Types: Geometry – Points



- Features that are too small to represent as lines or polygons (such as GPS observations).
- Can you think of any examples?

Image Source

21

Dataset Types: Geometry – Points Example

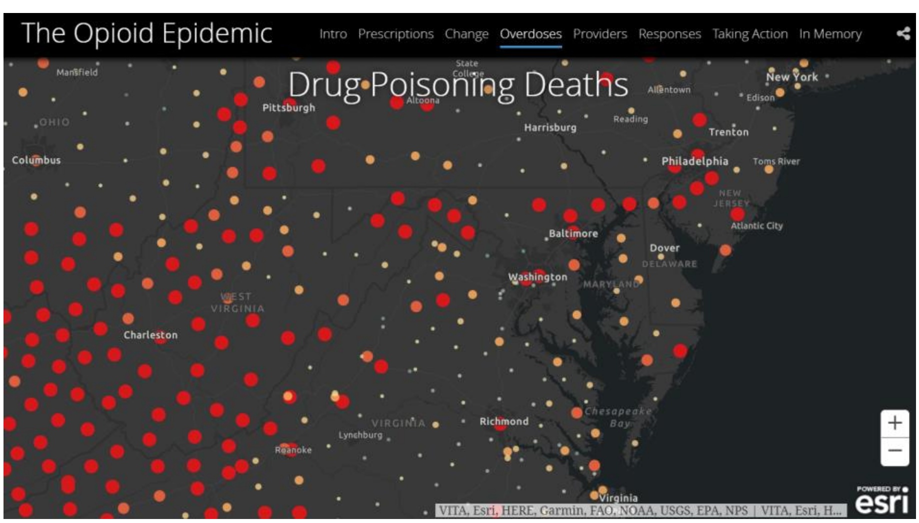
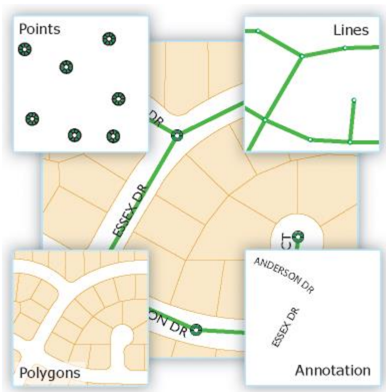


Image Source

22

Dataset Types: Geometry – Lines



- Represent the shape and location of geographic objects, such as street centerlines and streams
- Can you think of any examples?

Image Source

23

Dataset Types: Geometry – Lines Example

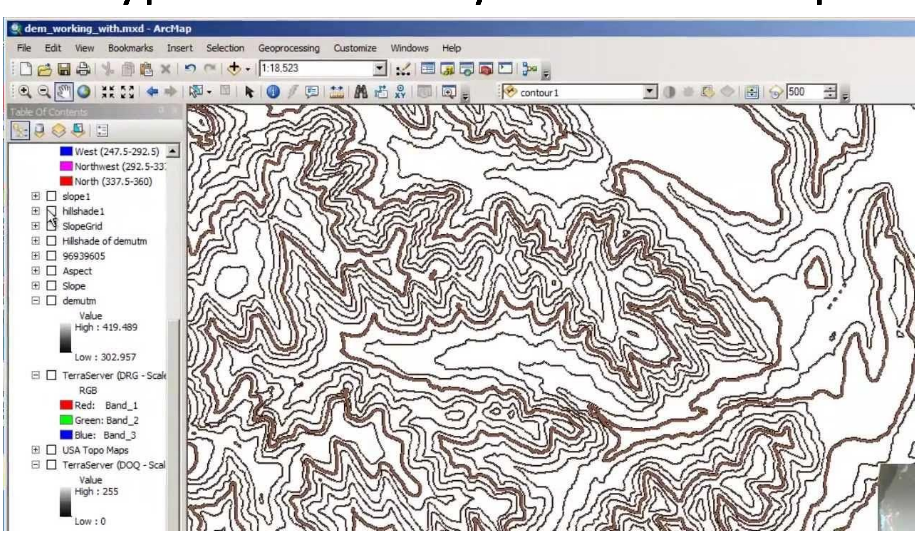
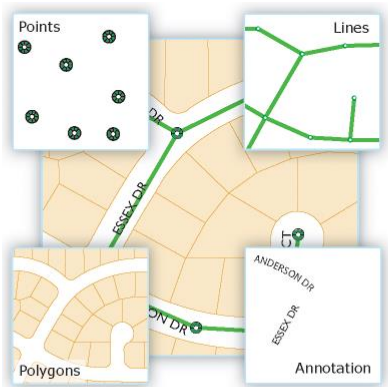


Image Source

24

Dataset Types: Geometry – Polygons



- A set of many-sided area features that represents the shape and location of homogeneous feature types
- Can you think of any examples?

Image Source

25

Dataset Types: Geometry – Polygons Example

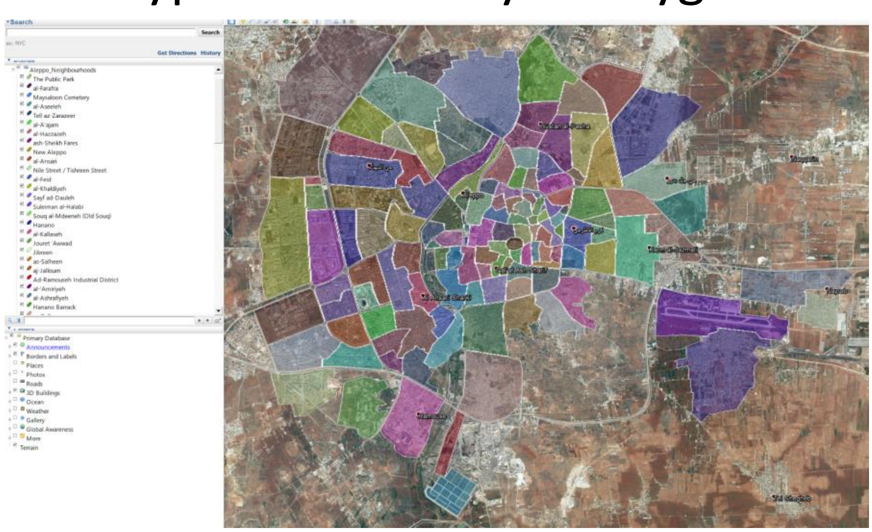
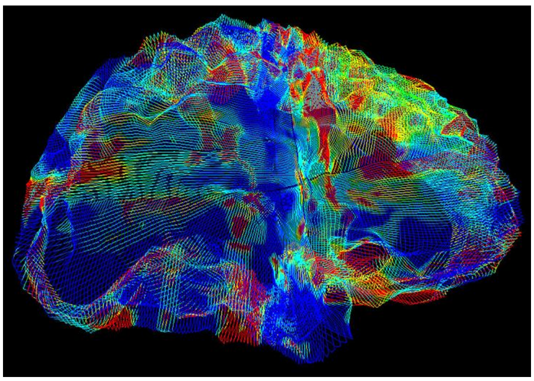


Image Source

26

Dataset Types: Geometry – Other



- Can also be used to map locations of pathologies in structures.
- Magnetic resonance imaging data of patients with Alzheimer's.

Image Source

27

Levels of Measurement

28

Levels of Measurement

- How something is measured is important
- For example, using a stopwatch to measure the time taken to respond to a stimulus
- This would not work for measuring attitude towards a political candidate
  - Rating scale is more appropriate

Osherson & Lane, 2007

29

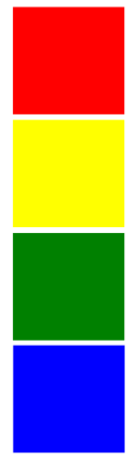
Major contributions of scale, by year

Author	Year	Title	Levels
Stevens	1946	Scales of Measurement	**nominal** **ordinal** **interval** ratio
Bertin	1967	Level of Organization of the Components	quantitative ordered quantitative quantitative
Harris	1966	Classification of Scales	category sequence quantitative quantitative
Munzner	2014	Visualization Principles	categorical/nominal ordinal quantitative quantitative
Börner	2014	Data Scale Types	nominal ordinal interval ratio

30

Levels of Measurement - Nominal

- Have already seen this with categorical variables
- Do not imply any ordering among the responses
- For example, when classifying people according to their favorite color
- Embody the lowest level of measurement



Osherson & Lane, 2007

31

Levels of Measurement - Ordinal

- Items in this scale are ordered
- Satisfaction with their microwave ovens:
  - Very dissatisfied
  - Somewhat dissatisfied
  - Somewhat satisfied
  - Very satisfied
- Adjacent scale values do not necessarily represent equal intervals on the underlying scale

Osherson & Lane, 2007

32

Levels of Measurement - Interval

- Intervals between points are consistent
- Difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees
- No true zero point

Osherson & Lane, 2007

33

Levels of Measurement - Ratio

- Like the three earlier scales rolled up in one:
  - Like a nominal scale, it provides a name or category for each object
  - Like an ordinal scale, the objects are ordered
  - Like an interval scale, the same difference at two places on the scale has the same meaning
  - Plus true zero
    - Zero money implies the absence of any money.

Osherson & Lane, 2007

34

Consequences of Level of Measurement

- Relationship between the variable's level of measurement and the statistics that can be meaningfully computed with that variable is important
- Could compute the mean of the codes but it would be meaningless

Color	Code
Blue	1
Red	2
Yellow	3
Green	4
Purple	5

Osherson & Lane, 2007

35

Consequences of Level of Measurement

- How about with an ordinal scale?
- Statisticians have debated for decades
- The prevailing opinion that for almost all practical situations, the mean of an ordinally-measured variable is a meaningful statistic
- Always consult with a statistician to make sure

Color	Code
Very dissatisfied	1
Somewhat dissatisfied	2
Somewhat satisfied	3
Very satisfied	4

Osherson & Lane, 2007

36

Question the Data

37

Question the Data: Questions

- What are the data requirements?
- What are the semantics of the data?
- What is the source of the data?
- Is the data accurate?
- What is the context of the data?
- What is the level of aggregation?

38

Question the Data Q1: Data Requirements

- "What data is required for the task at hand?"
- Analytics falls into two broad types:
  - Exploratory data analysis (EDA) does not begin with specific question but instead entails exploring data freely to get the lay of the land
  - Directed data analysis begins with one or more specific questions and then looks for answers
- Remaining questions (Q2 – Q5) assume that a data set has already been selected or provided

Few, 2019

39

Question the Data Q2: Data Semantics

- "What do the various fields of data mean?"
- Whenever we examine a data set, we must understand the semantics of the data:
  - Information you figure out from the data, versus the meanings that you must be told explicitly

Few, 2019

40

Why Data Semantics and Types Matter?

- Data visualization is driven by the kind of data that you have.
  - What information can you figure out from the data, versus the meanings that you must be told explicitly?
- Suppose you see the following data: 121/80, 121/75, 133/79, 101/87, 96/72
- It is hard to interpret the meaning of each number without more information

41

Type of Data

- Its structural or mathematical interpretation
- For example, the ratio of the systolic and diastolic blood pressure measurement

$$\frac{121}{80}$$

42

Semantics of Data

- Semantics of the data is its real-world meaning
- For example, how heart healthy you are

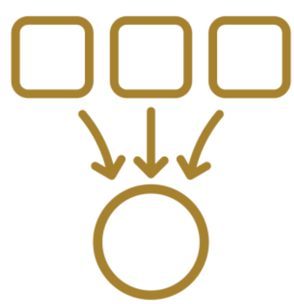
Blood Pressure Categories

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

43

Question the Data Q3: Data Source

- "What is the source of the data and is it credible?"
- Important to know the source of the data
- Important to know who produced the data:
  - Source of data might not be reliable or trustworthy
  - Data should be well-documented so that its origins can be easily traced



Few, 2019

44

Question the Data Q4: Data Accuracy

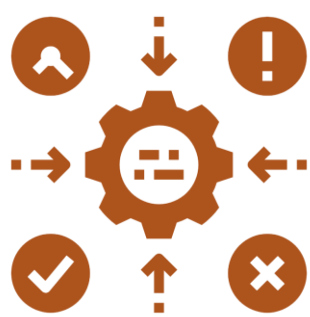
- “Is the data accurate?”
- Asking if the data is accurate is different from asking if its source is credible:
  - Credible sources can make errors
  - Unreliable sources can produce data that is accurate

Few, 2019

45

Question the Data Q5: Data Context

- "Have I taken all of the relevant context into account?“
- Nothing can be properly understood independent of its context
- Context is also learned from other sources, such as by reading an organization's publications, or by having conversations with staff



Few, 2019

46

Question the Data Q6: Data Aggregation

- "What is the level of data aggregation and is the statistic used to produce the aggregation appropriate?"
- Rare to visualize unaggregated data
- Different aggregation methods serve different purposes
- More about aggregation in an upcoming lecture

Few, 2019

47

Data Preparation

48

Data Dictionary

- Defines the characteristics of each variable
- If your data comes from a reputable source, it probably includes a data dictionary
- Figure is a sample from a data dictionary
- You will need to create one as part of your final project

Field/Variable	Definition
instant	Record index
dteday	Date
season	Season (1: winter, 2: spring, 3: summer, 4: fall)
Yr	Year (0: 2011, 1: 2012)
Mnth	Month (1 to 12)
Hr	Hour (0 to 23)
holiday	Whether day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
weekday	Day of the week
workingday	1: If day is neither weekend nor holiday, 0: otherwise
weathersit	1: Clear, Few clouds, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist; 3: Light snow, Light rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds; 4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
Temp	Normalized temperature in Fahrenheit
Atemp	Normalized feeling temperature in Fahrenheit
Hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	Count of casual users
registered	Count of registered users
Cnt	Count of total rental bikes including both casual and registered

49

Read Me Files

- Provides information about a data file and is intended to help ensure that the data can be correctly interpreted
- Standards-based metadata is generally preferable, if it exists
- Cornell Research Data Management Service Group has a very good template.

````
# Foobar

Foobar is a Python library for dealing with word pluralization.

## Installation

Use the package manager [pip](https://pip.pypa.io/en/stable/) to install
foobar.

```bash
pip install foobar
```

## Usage

```python
import foobar

foobar.pluralize('word') # returns 'words'
foobar.pluralize('goose') # returns 'geese'
foobar.singularize('phenomena') # returns 'phenomenon'
```

## Contributing
Pull requests are welcome. For major changes, please open an issue first to
discuss what you would like to change.

Please make sure to update tests as appropriate.

## License
[MIT](https://choosealicense.com/licenses/mit/)
````

Image Source

50

Common Table Formats (Types)

- Universal formats are comma-separated values (.csv), text (.txt), and Excel (.xlsx) files
- Use a file type that can be easily imported into most software used for data visualization such as Tableau, R, or Excel
- Addition file formats, might include data from a database (e.g., MySQL), a stats packages (e.g., SAS and SPSS), or other web-based formats (e.g., HTML, JSON, and XML)

Loth, 2019

51

Common Table Formats (Structure)

- Wide tables:
  - Many columns
  - Often summary tables containing aggregated measures (such as pivot tables in Excel)
  - Preprocessing of the data may be necessary.
- Long tables
  - Most of the time without aggregations
  - Each row containing one data point

Loth, 2019

52

Crosstab Reports with Wide Tables

- Common mistake is attempting to connect to a fully formatted Excel report that already shows data aggregations
- In the long run it is not worth it to work with this type of data

Sum of Sales		Region				
Sub-Category	Years	Central	East	South	West	Grand Total
Accessories	2016	4438.97	6053.768	5595.29	8926.244	25014.272
	2017	7795.228	17911.436	4141.534	10675.762	40523.96
	2018	10802.214	6231.378	9379.844	15482.418	41895.854
	2019	10919.664	14836.79	8160.086	26029.692	59946.232
Accessories Total		33956.076	45033.372	27276.754	61114.116	167380.318
Appliances	2016	3659.205	5779.202	2119.722	3755.496	15313.625
	2017	4974.509	6691.252	3850.34	7725.188	23241.289
	2018	6015.011	9426.582	5607.47	5001.252	26050.315
	2019	8933.308	12291.43	7947.794	13754.4	42926.932
Appliances Total		23582.033	34188.466	19525.326	30236.336	107532.161
Art	2016	821.954	1290.202	566.132	3379.694	6057.982
	2017	1132.16	1707.366	1362.318	2034.99	6236.834
	2018	1519.95	1882.566	1438.27	1120.122	5960.908
	2019	2291.276	2605.63	1288.902	2677.26	8863.068
Art Total		5765.34	7485.764	4655.622	9212.066	27118.792
Tables Total		39154.971	39139.807	43916.192	84754.562	206965.532
Grand Total		501239.8908	678781.24	391721.905	725457.8245	2297200.86

Loth, 2019

53

Data in Long Format

- Better to work with unaggregated raw data

Order Date	Segment	Region	Sub-Category	Sales
43412	Consumer	South	Bookcases	261.96
43412	Consumer	South	Chairs	731.94
43263	Corporate	West	Labels	14.62
43019	Consumer	South	Tables	957.5775
43019	Consumer	South	Storage	22.368
42530	Consumer	West	Furnishings	48.86
42530	Consumer	West	Art	7.28
42530	Consumer	West	Phones	907.152
42530	Consumer	West	Binders	18.504

Loth, 2019

54

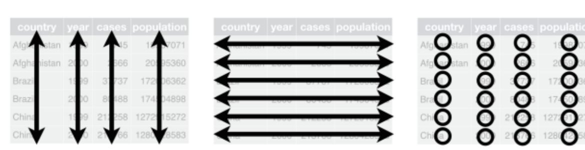
Data Wrangling

- Common cleaning steps might include:
  - Remove introductory text or metadata from the workbook
  - Pivot the data from the wide format to a long format
  - Ensure that numbers are formatted as such and not as text
  - Remove any empty rows
  - Make sure each column has a meaningful heading.

55

Tidy Data

- A consistent way to organize your data
- Getting your data into this format requires some up-front work, but that work pays off in the long term
- Format was developed to work with specific packages in R, it is a useful framework for most types of analysis



Wickham & Grolemund, 2014

56

Tidy Data: Each Column a Variable

- Do not created columns that contain two variables:
- For example, "male_treated" should be split into separate variables for sex and treatment status
- Store units in their own variable or in metadata, e.g., "3.4" instead of "3.4kg"

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

Not Tidy

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

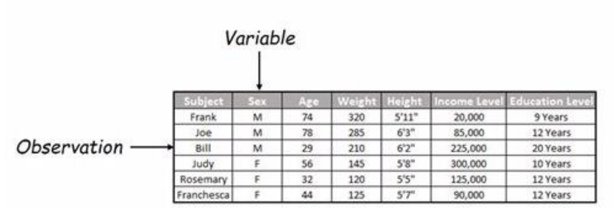
Tidy

Wilson et al., 2017

57

Tidy Data: Each Row an Observation

- For example, imagine 1 row per patient and then columns for measurements made for each patient.



Subject	Sex	Age	Weight	Height	Income Level	Education Level
Frank	M	74	320	5'11"	20,000	9 Years
Joe	M	78	285	6'3"	85,000	12 Years
Bill	M	29	210	6'2"	225,000	20 Years
Judy	F	56	145	5'8"	300,000	10 Years
Rosemary	F	32	120	5'5"	125,000	12 Years
Franchesca	F	44	125	5'7"	90,000	12 Years

Wilson et al., 2017

58

Data Check-list

- A Working with data check-list is loaded this checklist to Blackboard and I expect you to use it for your final project:
  - Defining the Problem
  - Locating and Retrieving Data
  - Data Preparation
  - Data Exploration

Rowell, Betzendahl, & Brown, 2020

59

Aggregation and Granularity

60

Reducing the Amount of Visible Data

- Reducing the amount of data shown in a view is an obvious way to reduce its visual complexity
- There are two primary methods for dimensions and measures:
  - Filtering, which eliminates elements
  - Aggregation combines many elements together
- The challenge is to minimize the chances that information important to the task is not lost in the transformation

Munzner, 2014

61

Aggregation

- Collecting values (individual numbers into a single value:
  - Summing all the sales for pumpkin spice lattes
  - Taking the average of all the temperature readings around Seattle on a given day

Benevento, Rowell, Steeger, Cutrell, & Morales, 2017a

62

Granularity

- Granularity is the ability to represent data, information, and knowledge at different levels of detail
- For example, in biology there are different levels of detail and hierarchical systems such as plant and animal taxonomies
- Understanding aggregation and granularity is critical concept because it affects how you build visualizations, how data is blended or joined, and how custom fields are created

Keet, 2013

63

Granularity: Examples

Sales by Sub-category, by Year

Sales by Sub-Category and Year

	Order Date			
Sub-Category	2011	2012	2013	2014
Accessories	113,456	172,398	209,895	253,488
Appliances	173,383	222,943	254,951	359,787
Art	64,139	82,358	98,007	127,588
Binders	86,999	93,418	121,075	160,420
Bookcases	259,396	317,953	376,026	513,197
Chairs	285,731	295,058	427,514	493,378
Copiers	216,368	327,169	415,515	550,385
Envelopes	27,987	38,014	50,805	54,099
Fasteners	13,609	19,478	21,597	28,559
Furnishings	63,934	81,804	111,820	128,020
Labels	13,616	15,518	18,381	25,889
Machines	160,546	159,859	198,376	260,279
Paper	42,666	51,512	70,513	79,601
Phones	337,282	364,016	453,519	552,006
Storage	205,627	228,556	309,476	383,427
Supplies	47,581	43,297	65,913	86,283
Tables	147,131	164,086	202,364	243,460

Sum of Sales broken down by Order Date Year vs. Sub-Category.

Sales by Sub-category, by Quarter

Sales by Sub-Category and Quarter

	Order Date			
Sub-Category	Q1	Q2	Q3	Q4
Accessories	103,268	169,360	221,154	255,455
Appliances	154,173	243,949	274,620	338,321
Art	57,593	88,214	111,390	114,896
Binders	70,050	98,262	138,437	155,163
Bookcases	232,430	303,435	439,693	491,015
Chairs	240,942	327,760	400,595	532,385
Copiers	238,260	346,126	398,852	526,199
Envelopes	26,691	40,732	48,483	54,998
Fasteners	12,051	21,241	21,742	28,207
Furnishings	62,269	87,252	104,558	131,499
Labels	10,388	17,576	21,396	24,044
Machines	143,705	176,739	189,517	269,099
Paper	35,910	58,645	66,172	83,565
Phones	231,883	396,700	491,681	586,560
Storage	191,502	257,438	310,444	367,702
Supplies	45,081	54,144	71,709	72,140
Tables	133,177	184,718	170,153	268,994

Sum of Sales broken down by Order Date Quarter vs. Sub-Category.

64

Journal Club

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. PLOS Computational Biology, 13(6).

65

Introduction

- Presents a set of good computing practices for any user
- Includes:
  - Data management
  - Collaborative programming
  - Organizing projects
  - Writing manuscripts
- Evidence-based from Carpentry workshops

THE CARPENTRIES

Wilson, 2017

66

Relevance to this lecture and class

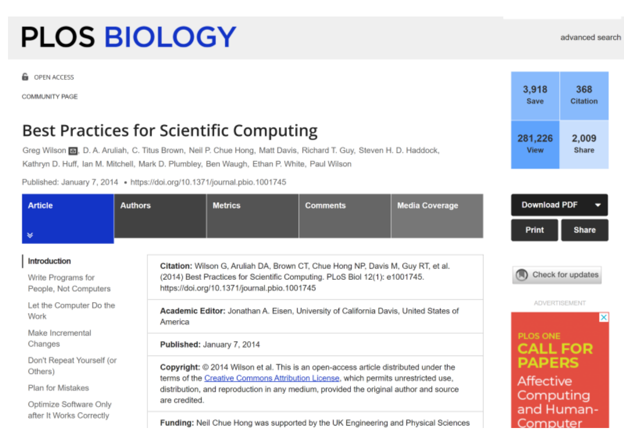
- 6 topic areas that are aligned to the work performed for the Final Project:
  - Effective data management
  - Effective writing, organizing, and sharing scripts
  - Uniform project organization
  - Learning about tools to help us write manuscripts

Wilson, 2017

67

Study and research question

- This is a retrospective study that uses data from a previous study
- Wilson, G. (2014). Best Practices for Scientific Computing. PLOS Biology, 12(1)
- Recommendations also influenced by other sources (see reference list)


PLOS BIOLOGY
advanced search

OPEN ACCESS

COMMUNITY PAGE

Best Practices for Scientific Computing

Greg Wilson, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

Published: January 7, 2014 • https://doi.org/10.1371/journal.pbio.1001745

3,918 Save | 368 Citation | 281,226 View | 2,009 Share

Article | Authors | Metrics | Comments | Media Coverage

Download PDF | Print | Share

Introduction
Write Programs for People, Not Computers
Let the Computer Do the Work
Make Incremental Changes
Don't Repeat Yourself (or Others)
Plan for Mistakes
Optimize Software Only after It Works Correctly

Citation: Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. (2014) Best Practices for Scientific Computing. PLoS Biol 12(1): e1001745. https://doi.org/10.1371/journal.pbio.1001745

Academic Editor: Jonathan A. Eisen, University of California Davis, United States of America

Published: January 7, 2014

Copyright: © 2014 Wilson et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Neil Chue Hong was supported by the UK Engineering and Physical Sciences

Check for updates

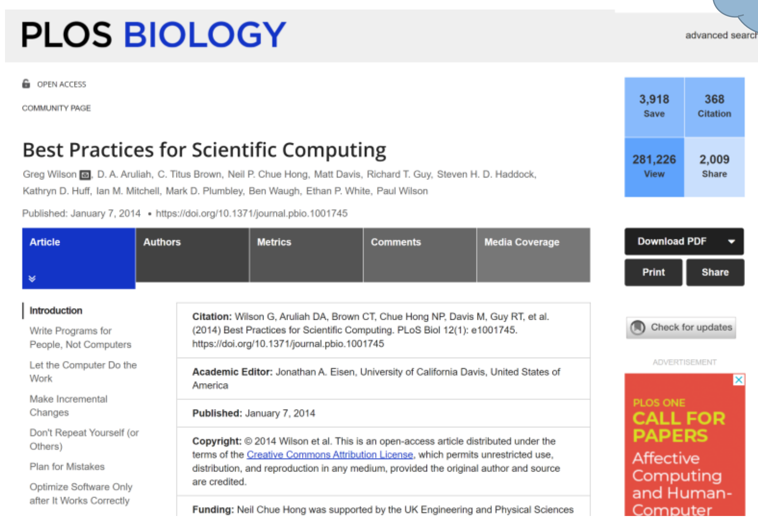
ADVERTISEMENT

PLOS ONE CALL FOR PAPERS Affective Computing and Human-Computer

68

Sidebar: Tip for Final Project

Click Me



69

Research Question

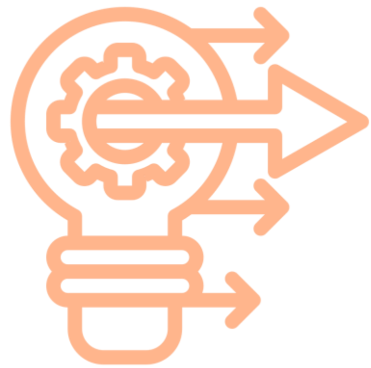
- Not an empirical study, so no H_o was developed
- See what I did there? For your presentations not all of the topic areas are relevant. However, these still need to be addressed during your presentation?
- How might be developed a H_o for a follow-up study?

Image source: @NASA

70

Methods

- Methods were not explicitly outlined in the paper, nor were there links to data in the supplementary data section



71

Critical Appraisal

- Based on my analysis of the paper, I thought it was balanced, and did not appear to be any conflicts of interest
- I would have like more information about data collecting and analysis

72

Summary of Results

- Practices are pragmatic
- Accessible to new users
- Increase productivity
- Supports reproducibility

73

How will this inform your practice

- I would like to go around the room and hear how each of you might use of the best practices:
  - Your final project
  - Your job
  - Other classes
- Let us start with the Zoom participants

74

Question for Discussion

- What was your general impression of the article?
- What was most inspiring, insightful, or surprising? What did you learn that you didn't expect to?
- Having read the article, what will you do differently the next time you analyze or visualize data?

Adapted from DataWrapper

75

Preview of lab

76

Lab Learning Objectives

- By the of this lab, learners should be able to:
  - Load external data from a .csv file into a data frame
  - Install and load packages.
  - Summarize the contents of a data frame
  - Understand the value of writing reproducible reports
  - Recognize and compile the basic components of an R Markdown file
  - Demonstrate the use of R code chunks, and understand their purpose, structure and option

77

R Packages

- Packages in R are basically sets of additional functions that let you do more stuff. The functions we learned about in the last lab come built into R.
- Before you use a package for the first time you need to install it on your machine, and then you should import it in every subsequent R session when you need it.

78

R Packages: Our Class Packages

- During the course we will need a number of R packages.
- Install these packages by opening RStudio and loading and running the following script:
- *lsc-563-install-packages.R*
- Using RStudio's graphical user interface by going to *Tools > Install Packages* and typing the names of the packages separated by a comma.

Tidyverse

- The major components of the tidyverse are:
  - Importing and formatting data
    - tibble: replaces data frames with tibbles
    - readr and readxl: facilitate data import and export
  - Wrangling data
    - dplyr and tidy: perform data manipulation
    - stringr: manipulate text strings
  - Visualizing data
    - ggplot2: data visualization library

Importing Data and Data frames

81

About the data

Column	Description
record_id	Unique id for the observation
month	month of observation
day	day of observation
year	year of observation
plot_id	ID of a particular plot
species_id	2-letter code
sex	sex of animal (“M”, “F”)
hindfoot_length	length of the hindfoot in mm
weight	weight of the animal in grams
genus	genus of animal
species	species of animal
taxon e.g.	Rodent, Reptile, Bird, Rabbit
plot_type	type of plot

- Data on species repartition and weight of animals caught in plots in our study area.
- The dataset is stored as a comma separated value (CSV) file.
- Each row holds information for a single animal.

82

Reading Data into R: Base R

These functions create a data structure known as a data frame:

- **read.table**: reads in tabular data from text file(s) where columns are separated by punctuation characters.
- **read.csv**: reads in comma separated data files.
- **read.delim**: reads in tab delimited data files.

We are going to use the R function **read.csv()** to load into memory the content of the CSV file as an object of class data.frame.

83

Reading in Data in R: Exercise

- Turn on the tidyverse using **library()** function
- We then need to import some data to work with, and save it to a object named: **surveys**
- Use: **read.csv**
- Dataset is: **combined.csv**

84

Notes About Reading in Data

- **read.csv** assumes that fields are delineated by commas, however, in several countries, the semicolon (;) is used as a field delineator.
- There is also the ***read.delim()*** for tab separated files. These functions are actually wrapper functions for ***read.table()***.
- As such, the surveys data above could have also been loaded by using read.table() with the separation argument as ",".

```
surveys2 <- read.table("raw_data/combined.csv", sep = ",", header = TRUE)
```

85

Data Frames (1)

- **Data frames** are the de facto data structure for most tabular data, and what we use for statistics and plotting.
- A data frame can be created by hand, but most commonly they are generated by the functions **read.csv()** or **read.table()**; in other words, when importing spreadsheets from your hard drive (or the web).

86

Data Frames (2)

- A **data frame** is the representation of data in the format of a table where the columns are vectors that all have the same length.
- Because columns are vectors, each column must contain a single type of data (e.g., characters, integers, factors).

1	"S"	TRUE
7	"A"	FALSE
3	"U"	TRUE
numeric	character	logical

87

Inspecting Data Frames (1)

- **dim()** - returns a vector with the number of rows in the first element, and the number of columns as the second element (the dimensions of the object)
- **nrow()** - returns the number of rows
- **ncol()** - returns the number of columns
- **head()** - shows the first 6 rows
- **tail()** - shows the last 6 rows

88

Inspecting Data Frames (2)

- **names()** – returns the column names (synonym of colnames() for data.frame objects)
- **rownames()** - returns the row names
- **str()** - structure of the object and information about the class, length and content of each column
- **summary()** - summary statistics for each column

89

Working with Data frames: Practice

90

Challenge 1

Based on the output of str(surveys), can you answer the following questions?

- What is the class of the object surveys?
- How many rows and how many columns are in this object?
- How many species have been recorded during these surveys?

91

About the Species Challenge (1)

Name	Genus	Species	Species ID	Count
Hispid pocket mouse*	Perognathus	Hispidus	PH	32

species versus species_id

Name	Genus	Species	Species ID	Count
Hispid cotton rat	Sigmodon	Hispidus	SH	147

** Chaetodipus hispidus is the valid name*

92

About the Species Challenge (2)

Species By Species (sp.)

Genus	Species	Species Id	
Chaetodipus	sp.	PX	6
Dipodomys	sp.	DX	40
Lizard	sp.	UL	4
Onychomys	sp.	OX	12
Pipilo	sp.	UP	8
Reithrodontomys	sp.	RX	2
Rodent	sp.	UR	10
Sparrow	sp.	US	4

93

Missing Data

94

Missing Data (1)

- Missing data in statistical programs can be represented in a number of different ways. For example, 999999, and period, or a dash.
- As R was designed to analyze datasets, it includes the concept of missing data (which is uncommon in other programming languages).
- Missing data are represented in vectors as *NA*, which R sees as any other value.

95

Missing Data (2)

- Most functions will return *NA* if the data you are working with include missing values.
- The simplest method for dealing with missing data is to add the argument **na.rm=TRUE** to calculate the result while ***ignoring the missing values***.
- There are a couple of common methods for detecting missing data:
  - *is.na()* function for checking for missing data
  - *na.omit()*
  - *complete.cases()*

96

Missing Data (3)

- Missing data becomes a problem when you try to compute summary statistics on a column with missing data.
- Mean will return a value of NA if even a single element is NA.
- This can be fixed by using the *na.rm = TRUE* command, which is added to the mean function.
- na.rm is a logical value indicating whether NA values should be stripped before the computation proceeds (R Core Team, 2017).
- Using not(!), you can extract those elements which are not missing values.

97

Exporting Data

98

Saving data from R (1)

- Similar to the *read_csv()* function used for reading CSV files into R, there is a *write_csv()* function that generates CSV files from data frames.
- We do not want to write "generated" datasets in the same directory as our raw data.
- Therefore, it is always a good practice to keep the raw data separate.
- Meaning that the **raw_data** folder should only contain the raw, unaltered data.

99

Saving data from R (2)

- Before using *write.csv()*, we need to make sure that we have a **data_output** folder in our project.
- We are now ready to save our data as a CSV file in our **data_output** folder.
- **write_csv**(x, "data_output/xxx.csv")

100

Bibliography

- Anscombe, F. J. (1973). Graphs in Statistical Analysis. American Statistician, 27(1), 17-21. doi:Doi 10.2307/2682899
- Benevento, D., Rowell, K., Steeger, J., Cutrell, A., & Morales, M. (2017). Tableau Desktop Interface and Navigation. In The Best Boring Book Ever™ of Tableau for Healthcare.
- Börner, K. (2015). Data Scale Types. In Atlas of knowledge : anyone can map (pp. 28-29). Cambridge, Massachusetts: The MIT Press.
- Börner, K., Sanyal, S., & Vespignani, A. (2007). Network Science. Annual Review of Information Science & Technology, 41.
- Cornell University. (2019). Guide to writing "readme" style metadata | Research Data Management Service Group. Retrieved from https://data.research.cornell.edu/content/readme

101

Bibliography

- Few, S. (2019). Question the Data. In The Data Loom (pp. 67-81). Burlingame, California: Analytics Press.
- Gehlenborg, N., & Wong, B. (2012). Networks. Nature Methods, 9, 115. doi:10.1038/nmeth.1862
- Healy, K. (2018a). Getting Started. In Data visualization: a practical introduction (pp. 32-53). Princeton, NJ: Princeton University Press.
- Healy, K. (2018b). Look at the Data. In Data visualization: a practical introduction (pp. 1-31). Princeton, NJ: Princeton University Press.
- Keet, C. M. (2013). Granularity. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), Encyclopedia of Systems Biology (pp. 850-853). New York, NY: Springer New York.

102

Bibliography

- Ledolter, J. (2013). Network Data. In Data Mining and Business Analytics with R (pp. 272-292).
- Loth, A. (2019). Introduction and Getting Started with Tableau. In Visual Analytics with Tableau (pp. 1-25). Indianapolis, IN: Wiley.
- Munzner, T. (2014a). Reduce Items and Attributes. In Visualization analysis and design (pp. 299-321). Boca Raton: CRC Press, Taylor & Francis Group.
- Munzner, T. (2014b). What: Data Abstraction. In Visualization analysis and design (pp. 20-40). Boca Raton: CRC Press, Taylor & Francis Group.
- Myatt, G. J., & Johnson, W. P. (2014). Introduction. In G. J. a. J. Myatt, W. P. (Ed.), Making Sense of Data I.

103

Bibliography

- Rowell, K. L., Betzendahl, L., & Brown, C. (2020). Stop Hunting Unicorns and Start Building Teams and Know the Data. In Visualizing health and healthcare data : creating clear and compelling visualizations to "see how you're doing (pp. 9-36). Hoboken, NJ: John Wiley & Sons.
- Ware, C. (2013). Foundations for an Applied Science of Data Visualization. In Information Visualization: Perception for Design (3rd ed., pp. 1-30). Waltham, MA: Morgan Kaufmann (Elsevier).
- Wickham, H., & Grolemund, G. (2014). Tidy Data. In R for Data Science: Import, Tidy, Transform, Visualize, and Model Data (pp. 147-170).
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. PLOS Computational Biology, 13(6), e1005510-e1005510. doi:10.1371/journal.pcbi.1005510

104