Week 02

Aggregating & Summarization

INFO 3402: Information Exposition

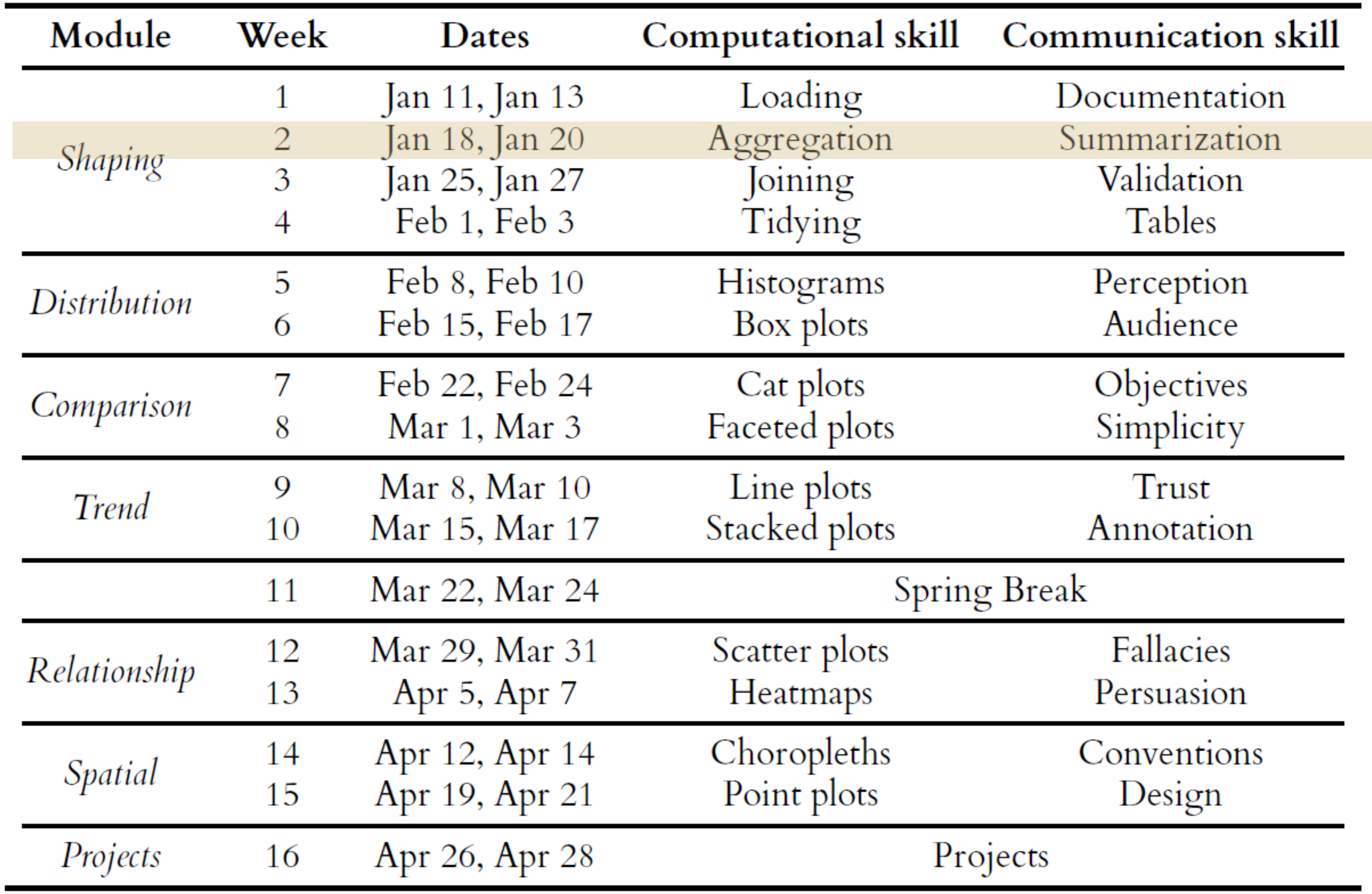
Brian Keegan, Ph.D.
Assistant Professor, Information Science
brian.keegan@colorado.edu

University of Colorado
Boulder

Agenda

- 00:00 – 00:15 → Slides
- 00:15 – 01:00 → Notebook
- 01:00 – 01:15 → Exercises

Course Overview



Module	Week	Dates	Computational skill	Communication skill
Shaping	1	Jan 11, Jan 13	Loading	Documentation
	2	Jan 18, Jan 20	Aggregation	Summarization
	3	Jan 25, Jan 27	Joining	Validation
	4	Feb 1, Feb 3	Tidying	Tables
Distribution	5	Feb 8, Feb 10	Histograms	Perception
	6	Feb 15, Feb 17	Box plots	Audience
Comparison	7	Feb 22, Feb 24	Cat plots	Objectives
	8	Mar 1, Mar 3	Faceted plots	Simplicity
Trend	9	Mar 8, Mar 10	Line plots	Trust
	10	Mar 15, Mar 17	Stacked plots	Annotation
	11	Mar 22, Mar 24	Spring Break	
Relationship	12	Mar 29, Mar 31	Scatter plots	Fallacies
	13	Apr 5, Apr 7	Heatmaps	Persuasion
Spatial	14	Apr 12, Apr 14	Choropleths	Conventions
	15	Apr 19, Apr 21	Point plots	Design
Projects	16	Apr 26, Apr 28	Projects	

Weekly overview

- Notebook exercises (ungraded) – Assigned Tuesdays and reviewed Thursdays
- Add a Visualization Critique (ungraded) – Discussed in class on Thursdays
- Weekly Assignment (graded, 2%) – Assigned Tuesdays and due Fridays
- Weekly Quiz (graded, 1%) – End of class on Thursdays (12pm)

Types of data

Types of variables - Categorical

- **Dichotomous** variables have only two values
  - Examples: Booleans (True/False), Sex (Male/Female), Possession (Yes/No)
  - Finite → exactly two possible values
  - Unordered → no meaningful greater or less than relationship
- **Nominal** variables are finite and unordered
  - Examples: Race (White/Black/Asian), Marital Status (Married/Single/Divorced/Widowed)
  - Finite → all the different values are defined and countable (ideally, but not necessarily, small)
  - Unordered → the values have no meaningful greater or less than relationships
- **Ordinal** variables are finite and ordered
  - Examples: Likert (Support/Oppose), education (HS/College/Grad School), income (Low/Medium/High)
  - Finite → values capture distinct ranges
  - Ordered → some values are greater or less than others
- Categorical variables are repeated in data → good to use as a group, rarely used to aggregate

Types of variables - Continuous

- **Continuous** variables are infinite and ordered
  - Examples: Age, weight, height
  - Uncountable → Infinite number of values possible between two numbers
  - Ordered → values are greater or less than other values, can be sorted
  - Precision is important (16.00007g is technically but not meaningfully greater than 16g)
  - Parametric statistics are good aggregation functions (mean, standard deviation)
- **Discrete** variables are finite and ordered
  - Examples: GPA, counts, points, IDs
  - Countable → values are defined and cannot take on intermediate values (cannot own 1.5 apples)
  - Ordered → values are greater or less than other values, can be sorted
  - Better to use non-parametric statistics to aggregate (median, mode, counts)
- Repeated values possible but rare → they should *not* be used to group, but are *great* to aggregate

Combining data types

- Groupby-aggregation requires a combination of categorical and continuous data
- You'll often-to-always use *categorical* data to create a groupby object
  - Continuous data rarely has identically repeated values
- You'll generally-to-often use *continuous* data to aggregate
  - Many aggregation functions do not apply to categorical data (mean, sum, *etc*.)
- **General rule**: Groupby categorical variables and aggregate on continuous variables

Summarizng

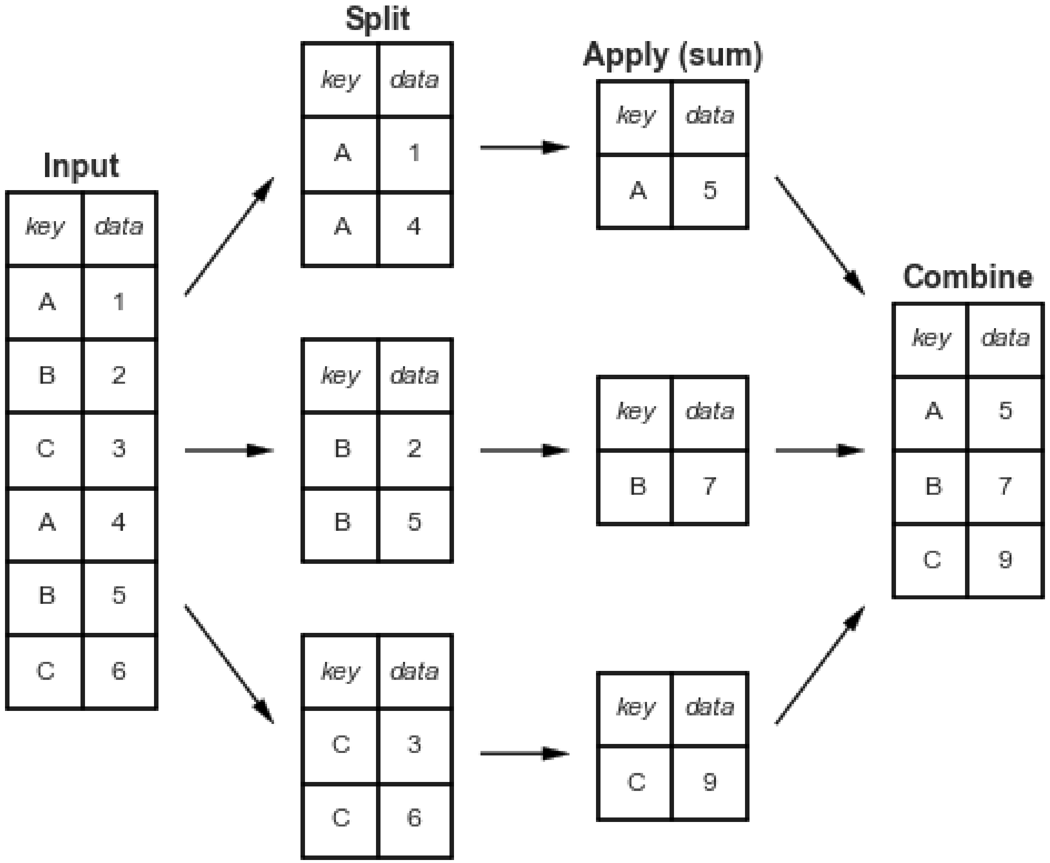
Classic summarizing functions

- We need to apply a function to summarize the values present in each groupby baby DataFrame

- **Central tendency**: mean, median, mode
- **Extent**: len, size, min, max
- **Total**: sum, product
- **Position**: first, last, idxmin, idxmax
- **Distribution**: value_counts

- More advanced summaries can be applied with lambda functions, custom functions, *etc*.

Split-Apply-Combine



Input

key	data
A	1
B	2
C	3
A	4
B	5
C	6

Split

key	data
A	1
A	4

key	data
B	2
B	5

key	data
C	3
C	6

Apply (sum)

key	data
A	5

key	data
B	7

key	data
C	9

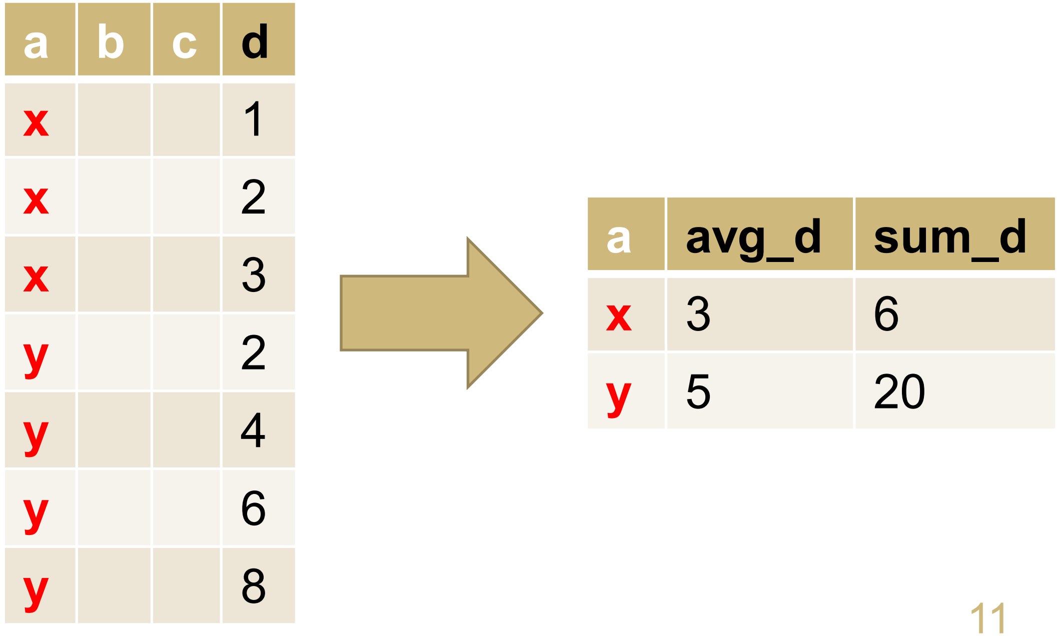
Combine

key	data
A	5
B	7
C	9

Groupby-Aggregation

- Many forms of data have repeated values that we want to summarize
  - States over time, schools in a district, events on a date, actions by an account
- Think about this as combining **rows** together

```
df.groupby(‘a’).agg({‘d’:[‘mean’,’sum’]})
```



a	b	c	d
x			1
x			2
x			3
y			2
y			4
y			6
y			8

a	avg_d	sum_d
x	3	6
y	5	20

Notebook Time!

- Download the “Week 02 – Lecture.ipynb” and “CDC_deaths_2014_2022.csv” files
  - Please create a dedicated folder for class instead of keeping everything in Downloads, Desktop, *etc*.
  - Put both these files in the same folder
- Open the “Week 02 – Lecture.ipynb” notebook file
  - From Anaconda Prompt (Windows) or Terminal (Mac), navigate to class folder
  - Launch Jupyter Notebook: `jupyter notebook`

- Work on Exercises 1-5, practice consulting documentation and other resources
- No grading on exercises, we’ll cover solutions on Thursday

Weekly Assignment & Next Class

Weekly Assignment 02

- **Skills**: Groupby-aggregation and pivoting data with continuous and categorical variables
- **Data**: Longitudinal time use survey data

- Due Friday before midnight on Canvas
  - Save an HTML version of your notebook with all output present
    - File > Download as > HTML (.html)
  - Upload the HTML file to Canvas

Next Class

- Review concepts and exercises from last class
- Critique a data narrative or visualization
  - Post a link and a few sentences about a data visualization on Canvas discussion (ungraded/optional)
- Time to work on Weekly Assignment
  - Weekly Assignment due on Friday by submitting HTML notebook to Canvas before midnight
- Weekly quiz at the end of class