School of Economics and Management

Information Science III

2. Data

Yuki Yanai

🌐 https://yukiyanai.github.io

✉ yanai.yuki@kochi-tech.ac.jp

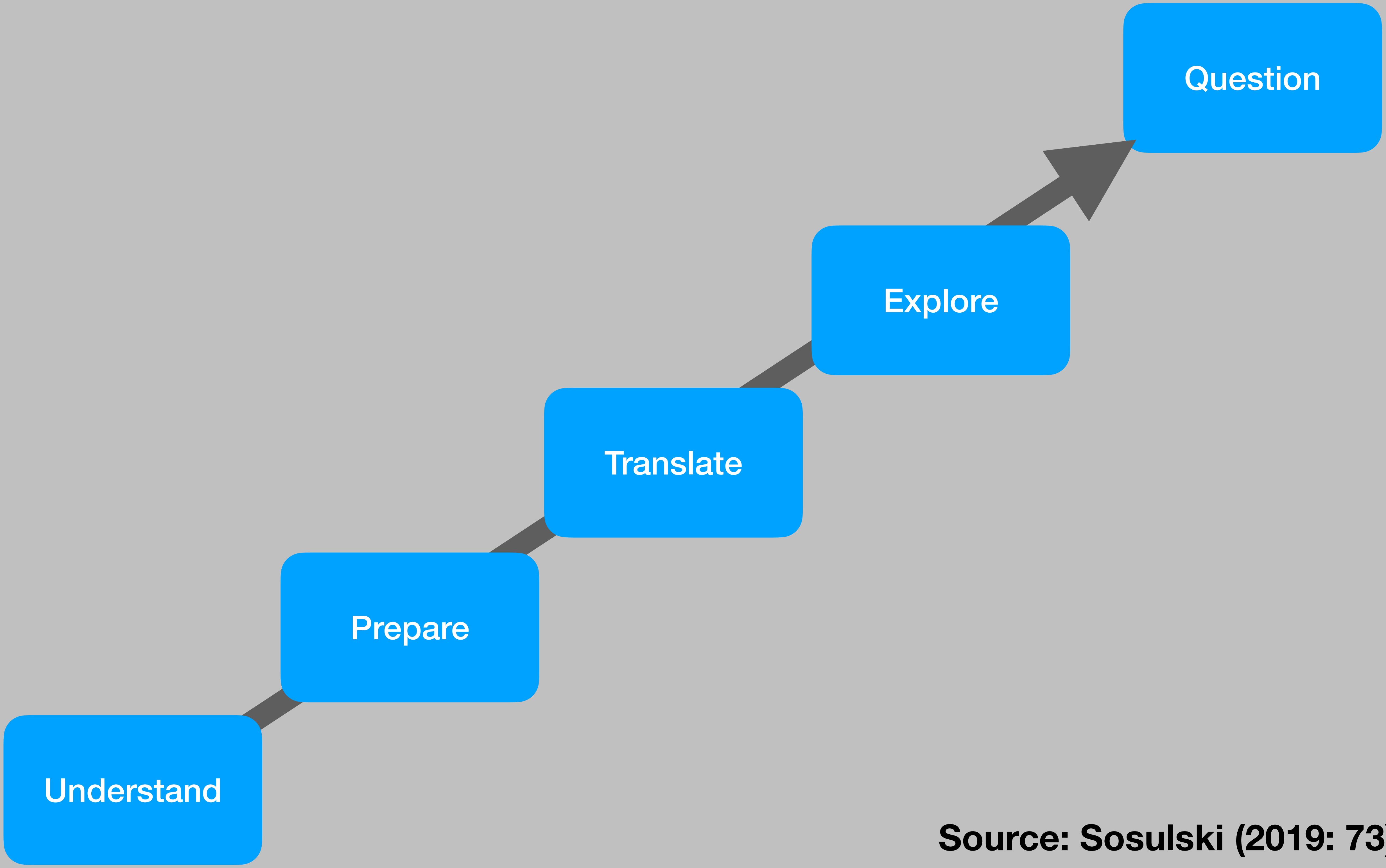
Today's Goals

- To understand:
  - What data are
  - How we extract information from data
  - How we should transform data before visualizing them

Data Handling

Preparation for Visualization

Extract Information from Data



Source: Sosulski (2019: 73)

Get Data You Are Interested In

- Governments and public organizations' data
  - e-Stat
  - World Bank Open Data
  - IMF Data
- Surveys
  - International Social Survey Programme
  - World Values Survey
- Many other data are available!

Data Format

- General purpose data format
  - .csv (comma separated values)
  - .tsv (tab separated values)
  - .txt (text data; table format data)
- Data for specific applications
  - .xlsx or .xls (Excel)
  - .Rds, .RData (R)
  - .dta (Stata) or .sav (SPSS) or .sas (SAS)
- Web-format data
  - .html, .xlm, .json

Understand Your Data

- View the dataset
  - Open the dataset with a spreadsheet application
    - E.g., LibreOffice Calc, Microsoft Excel
  - Read the dataset with R
- Read the codebook (dictionary) of the data
  - What does each variable measures?
  - What do values of each variable represent?
  - How were the data collected?

Example: Bike Sharing Data

- Some questions about bike sharing
  - What time of year is most popular for bike rentals?
  - What's the most popular day of the week for bike rentals?
  - What's the frequency of use for the average user?
  - What are the most and least congested bike stations?
- Get the data: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset (or google "bike sharing data")

Understand Bike Sharing Data

- What are in rows?
  - How many rows? (Or what is the sample size?)
- What are in columns?
  - What variables does the file contain?
- What does each variable measure?
  - Read the codebook, dictionary, or Readme

Transform Data

- Most of the times, you cannot analyze data you got as they are
- You need to transform the data set somehow
  - Rename variables
  - Change the type of variables (e.g., character to factor)
  - Decide how to deal with missing values
  - Scale values (e.g., standardize, take natural log)
  - Transform wide data into long data
  - Aggregate values by group

Tidy Data

(Review)

How Should We Prepare a Dataset?

- To analyze data with R, we need a dataset in a nice format
  - Something that we can easily handle
  - One answer: tidy data

Tidy Data

- Proposed by Hadley Wickham
- Tidy data: structure and meaning matches
- Non tidy data: messy data
- We want to prepare tidy data for our data visualization and analysis

Four Conditions of Tidy Data

1. Each variable is a column
2. Each observation is a row
3. Each type of observational unit is a table
4. Each value is a cell

Weather in 3 Cities: Messy Data

City	6	12	18
Kochi	Sunny	Sunny	Cloudy
Tokyo	Cloudy	Rainy	Rainy
Osaka	Rainy	Sunny	Sunny

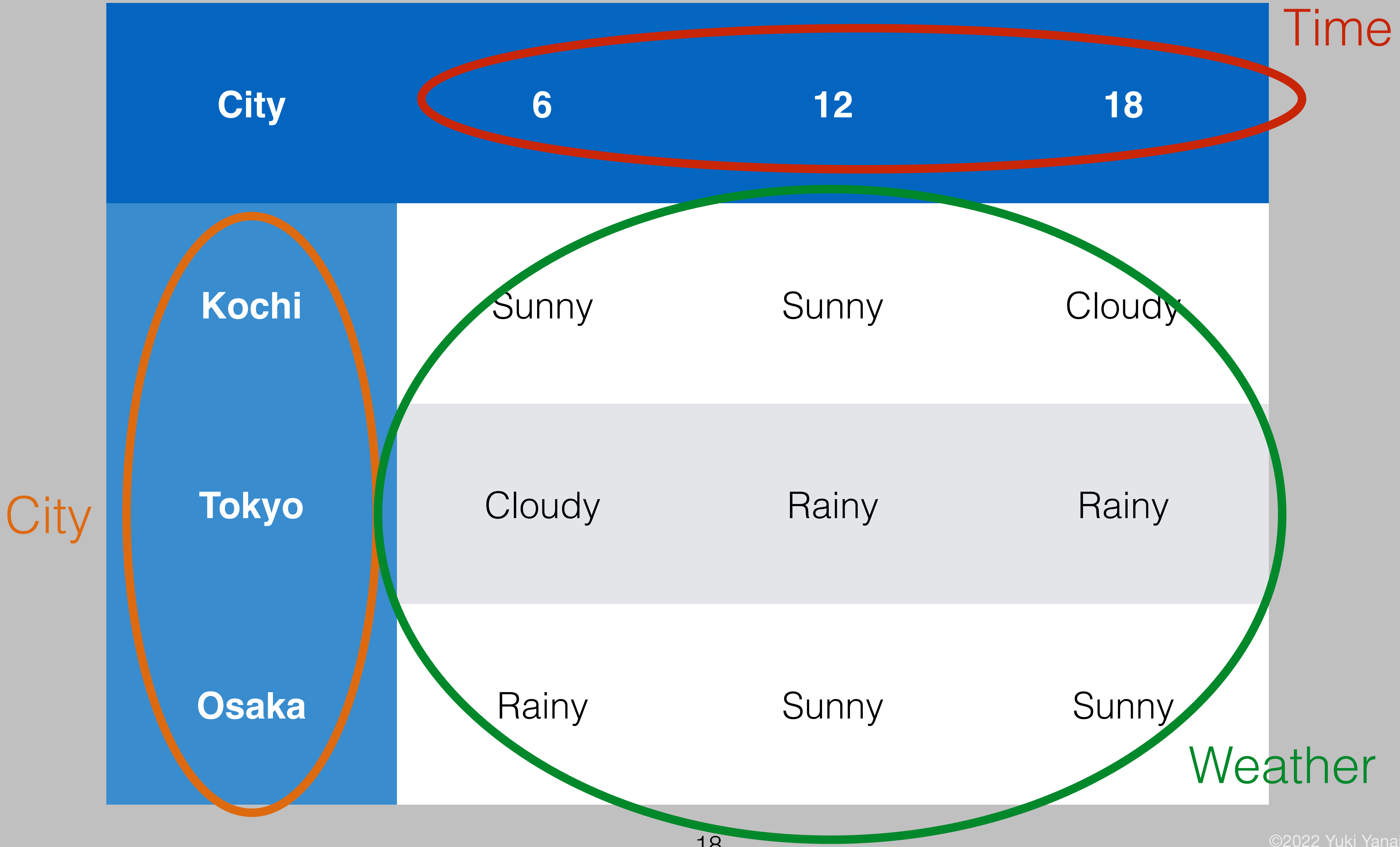
Weather in 3 Cities: Tidy Data

City	Time	Weather
Kochi	6	Sunny
Kochi	12	Sunny
Kochi	18	Cloudy
Tokyo	6	Cloudy
Tokyo	12	Rainy
Tokyo	18	Rainy
Osaka	6	Rainy
Osaka	12	Sunny
Osaka	18	Sunny

Tidy vs. Messy Data

- Tidy data are not always better than messy data
  - To human eyes, messy data might look nicer: our example of weather
- However, for data analysis, tidy data is better, because it is easier to handle them than messy data

Variables and Columns in Messy Data

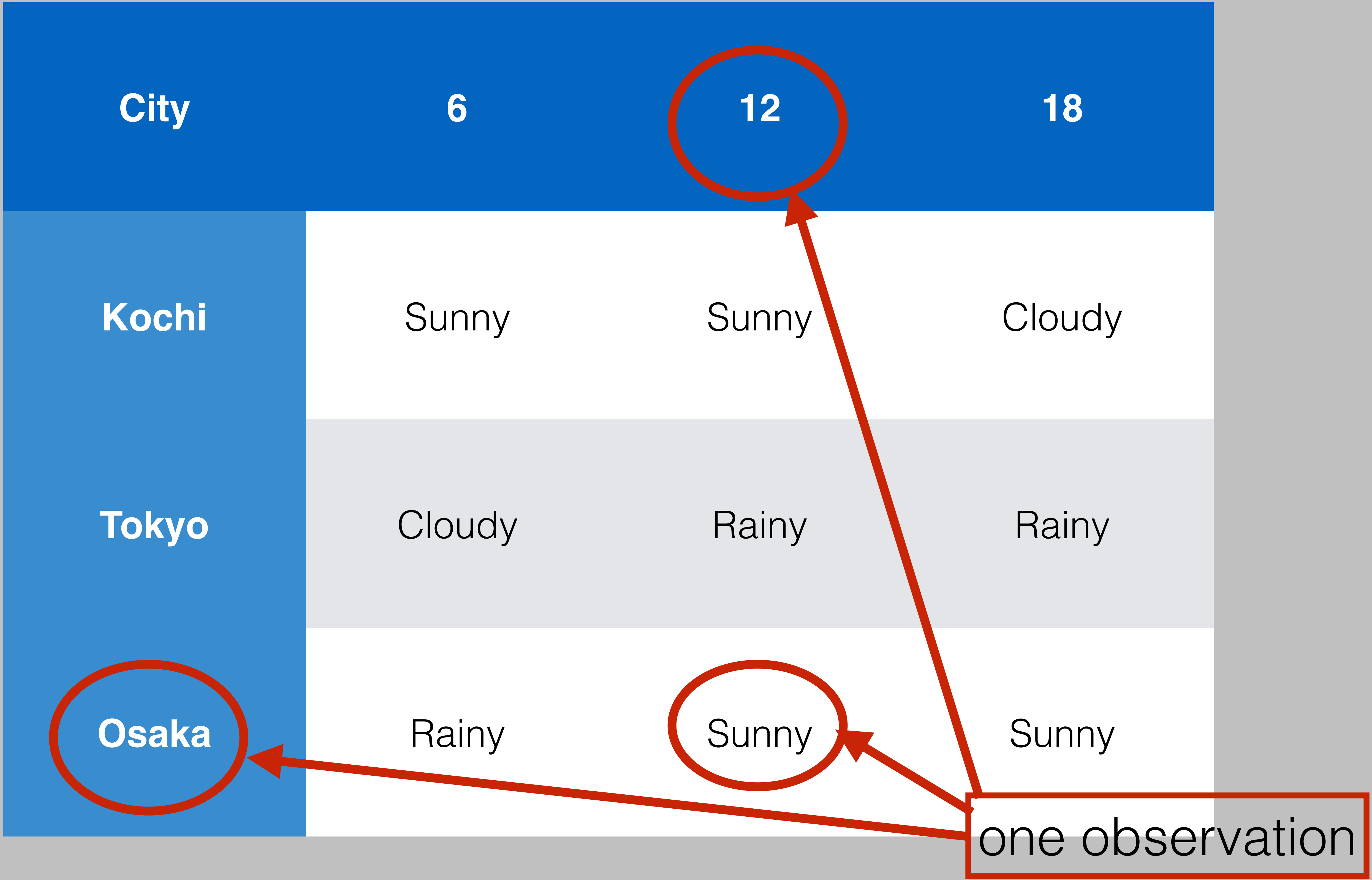


City	6	12	18
Kochi	Sunny	Sunny	Cloudy
Tokyo	Cloudy	Rainy	Rainy
Osaka	Rainy	Sunny	Sunny

Variables and Columns in Tidy Data

City	Time	Weather
Kochi	6	Sunny
Kochi	12	Sunny
Kochi	18	Cloudy
Tokyo	6	Cloudy
Tokyo	12	Rainy
Tokyo	18	Rainy
Osaka	6	Rainy
Osaka	12	Sunny
Osaka	18	Sunny

Observations and Rows in Messy Data



City	6	12	18
Kochi	Sunny	Sunny	Cloudy
Tokyo	Cloudy	Rainy	Rainy
Osaka	Rainy	Sunny	Sunny

one observation

Observations and Rows in Tidy Data

City	Time	Weather
Kochi	6	Sunny
Kochi	12	Sunny
Kochi	18	Cloudy
Tokyo	6	Cloudy
Tokyo	12	Rainy
Tokyo	18	Rainy
Osaka	6	Rainy
Osaka	12	Sunny
Osaka	18	Sunny

one observation

Each Type of Observational Unit is a Table

- In a single table (or dataset), you have only one type of observational unit
    - E.g. Each observation is an individual person, or each observation is a country

Messy Data with Multiple Types of Observational Unit

Country	Presidential?	City	Population (million)
Japan	No	Tokyo	9.4
Japan	No	Osaka	2.7
Japan	No	Nagoya	2.3
USA	Yes	New York	8.5
USA	Yes	Chicago	2.7
USA	Yes	Los Angles	3.9

Observational Unit:
Country

Observational Unit:
City

Tidy Data with One Types of Observational Unit

Key to connects two tables

City	Population (million)	Country
Tokyo	9.4	Japan
Osaka	2.7	Japan
Nagoya	2.3	Japan
New York	8.5	USA
Chicago	2.7	USA
Los Angles	3.9	USA

Observational Unit:
City

Country	Presidential
Japan	No
USA	Yes

Observational Unit:
Country

Each Value Is a Cell

Tidy Data

City	Time	Weather
Kochi	6	Sunny
Kochi	12	Sunny
Kochi	18	Cloudy
Tokyo	6	Cloudy
Tokyo	12	Rainy
Tokyo	18	Rainy
Osaka	6	Rainy
Osaka	12	Sunny
Osaka	18	Sunny

Messy Data

City	Time	Weather
Kochi	6 & 12	Sunny
Kochi	18	Cloudy
Tokyo	6	Cloudy
Tokyo	12 & 18	Rainy
Osaka	6	Rainy
Osaka	12 & 18	Sunny

Structures and Meanings Should Match

- In tidy data
  - Column: a variable
  - Row: an observation
  - Cell: a value
  - Table: information of one type of observational unit
- We want to know meanings of relationship between variables
- When we analyze data, we write commands that unitize the structure of data

Exploring Data by Making Graphs

Statistics

- First step of exploring data: calculate statistics
  - Central tendency: mean, median, mode
  - Variability of the variable: variance, standard deviation, rages, IQR
  - More details: kurtosis, skewness, etc.
- Might need to transform data to calculate statistics by group
  - E.g., Bike rentals by month

A Variety of Plots

- There exist a lot of different types of plots: E.g.,
  - Bar charts
  - Histograms / density plots
  - Box[-and-whisker] plots / violin plots
  - Scatter plots
  - Line plots
- Need to choose the best one for your purpose

Exercises

- Show the following by both statistics and graphs
  - Frequency of bike rentals in 2012
  - Bike rentals by month in 2012
  - Relationship between temperature and bike rentals
  - Differences in bike rentals by month between 2011 and 2012
- First assignment
  - Make a single pdf file containing the answers to the above questions
    - Filename: `info3_YourName_hw01.pdf` (YourName should be your name: e.g., `info3_YukiYanai_hw01.pdf`)
  - **Deadline: 6 pm on Thursday, Oct 13, 2022**
  - **Submit the file at KULMS**

Next class

3. Good Visualization