国立情報学研究所
National Institute of Informatics

Acquiring Practical Skills of Data Science

WEEK7

~ Data Visualization ~

Schedule

01 (4/15)	:	Introduction
02 (4/22)	:	Software
03 (5/13) & 04 (5/20)	:	Programming
05 (5/27) & 06 (6/03)	:	Data Acquision & Construction
07 (6/10)	:	Data Visualization
08 (6/17)	:	Presentation
09 (6/24) & 10 (7/01)	:	Data Analysis
11 (7/08) & 12 (7/22)	:	Simulation
13 (7/29)	:	Data Science Literacy
14 (8/05) & 15 (8/06)	:	Presentation

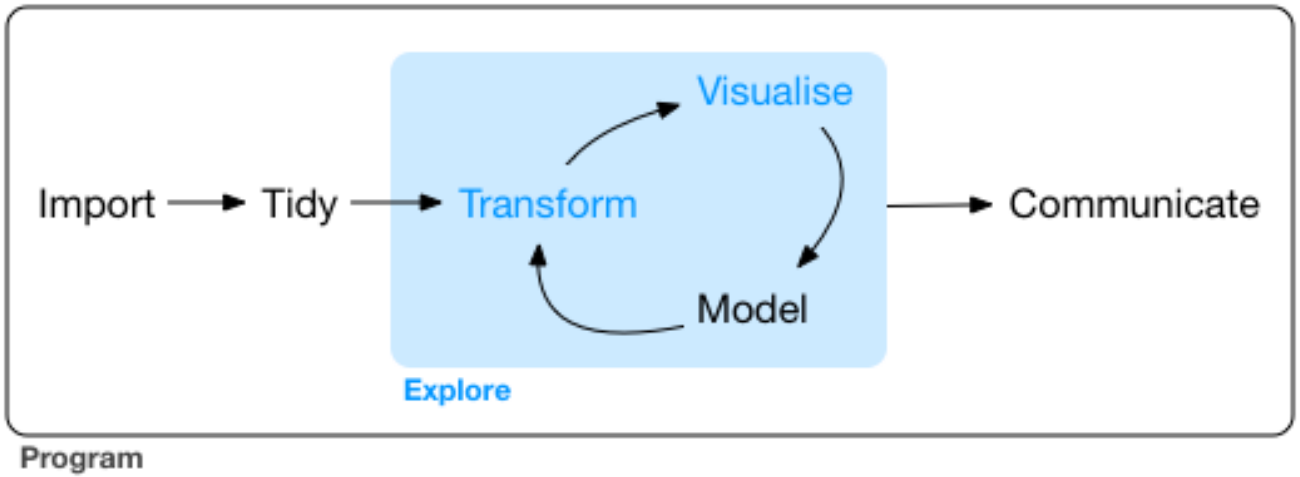
Week 07 (6/10): Data Visualization

- Goal for week07： Become able to build a basic data visualization

- Topics
  - What is data visualization?
  - Building plots with ggplot2
    - Creating a plot in major visualization types
    - Exporting a plot to a file

- Environment
  - Programming Language: R language
  - Platform: Jupyter Hub
  - Form : Coding for Exercise Tasks on Jupyter Hub

Exploratory data analysis



From Hadley Wickham & Garret Grolemund, R for Data Science, O'Reilly

Exploratory Data Analysis Checklist

1. Formulate your question
2. Read in your data
3. Check the packaging
4. Run str()
5. Look at the top and the bottom of your data
6. Check your “n”s
7. Validate with at least on external data source
8. Try the easy solution first
9. Challenge your solution
10. Follow up

1. Formulate your question

- A sharper question or hypothesis is easier
  - Eliminate variables that are not relevant to the question
- Q. Are air pollution levels higher on the east coast than on the west coast?
  - <= all pollutants across entire east and west
- Q. Are hourly ozone levels on average higher in New York City than they are in Los Angeles?
  - <= single pollutant in two cities

2. Read in your data

- Data is sometimes/always messy
- Cleaning up a dataset
- ”Tidy data”

3-5. Check data

- Check the packaging
  - Find warnings or errors when reading
  - Check the number of rows and columns after reading
    - nrow(data), ncol(data)
- Check the content briefly
  - str(data)
- Look at the top and the bottom
  - head(data), tail(data)

6-7. Check data

- Check your “n”s
  - Identify some landmarks that can be used to check
  - ”Does it include expected Date and Time properly?”
  - “Does it cover all of states?”
- Validate with at least one external data source
  - Measurements: summary()
  - Distributions: quantile()
  - Units

8. Try the easy solution first

- Use simple measurements
- Use a portion of data
  - Top 10, Bottom 10, ...
- Use a group of data
  - By month, By year, By country
  - filter()

- 9. Challenge your solution
  - The easy solution is nice, but...
  - You should always think of ways to challenge results
  - A result by year was great
    - By month? Is there enough data?
- 10. Follow up questions
  - Do you have the right data?
  - Do you need other data?
  - Do you have the right question?

Principles of Analytic Graphics

- Show comparisons
- Show causality, mechanism, explanation, systematic structure
- Show multivariate data
- Integrate evidence
- Describe and document the evidence
- Content, content, content, ...

Materials

- https://github.com/fumi/DS2019_Week07

Week 08 (6/17): Presentation

- Report
  - Choose a topic you are interested in
  - Use techniques learned in Data Acquision & Construction - Data Visualization
  - Within 3000 English words
  - Due on 6/17
  - Hand in your report to: yamaji@nii.ac.jp
- Presentation
  - Make a presentation based on the report
  - Within 15 minutes