Topic 2: Exploratory data analysis

Details

Sonja Petrović

Created for ITMD/ITMS/STAT 514

Spring 2021.

Exploratory data analysis[1]

- Use visualisation and transformation to explore your data in a systematic way. EDA is an iterative cycle. You:
  - Generate questions about your data.
  - Search for answers by visualising, transforming, and modelling your data.
  - Use what you learn to refine your questions and/or generate new questions.
- *Not* a formal process with a strict set of rules; "a state of mind."
- EDA is an important part of any data analysis, even if the questions are handed to you on a platter, because you always need to investigate the quality of your data. Data cleaning is just one application of EDA: you ask questions about whether your data meets your expectations or not. To do data cleaning, you'll need to deploy all the tools of EDA:
  - visualisation,
  - transformation, and
  - modelling.

[1] https://r4ds.had.co.nz/exploratory-data-analysis.html

Questions

"There are no routine statistical questions, only questionable statistical routines." — Sir David Cox

Goal:

develop an understanding of your data.

- EDA is fundamentally a creative process: the key to asking *quality* questions is to generate a large *quantity* of questions.
- There isn't one rule, but here is a guideline:
  - What type of variation occurs within my variables?
  - What type of covariation occurs between my variables?
- Concepts to remember: variable, value, observation, tabular data.

Variation

Recall:

> *Variation is the tendency of the values of a variable to change from measurement to measurement.*

To understand, we try to visualize distributions.

- Bar charts (type of random variable:

Variation

Recall:

> *Variation is the tendency of the values of a variable to change from measurement to measurement.*

To understand, we try to visualize distributions.

- Bar charts (type of random variable: discrete/categorical)
- Histograms (type of random variable:

Variation

Recall:

> *Variation is the tendency of the values of a variable to change from measurement to measurement.*

To understand, we try to visualize distributions.

- Bar charts (type of random variable: discrete/categorical)
- Histograms (type of random variable: continuous)
  A quick summary: https://r4ds.had.co.nz/exploratory-data-analysis.html#variation

Visualizations

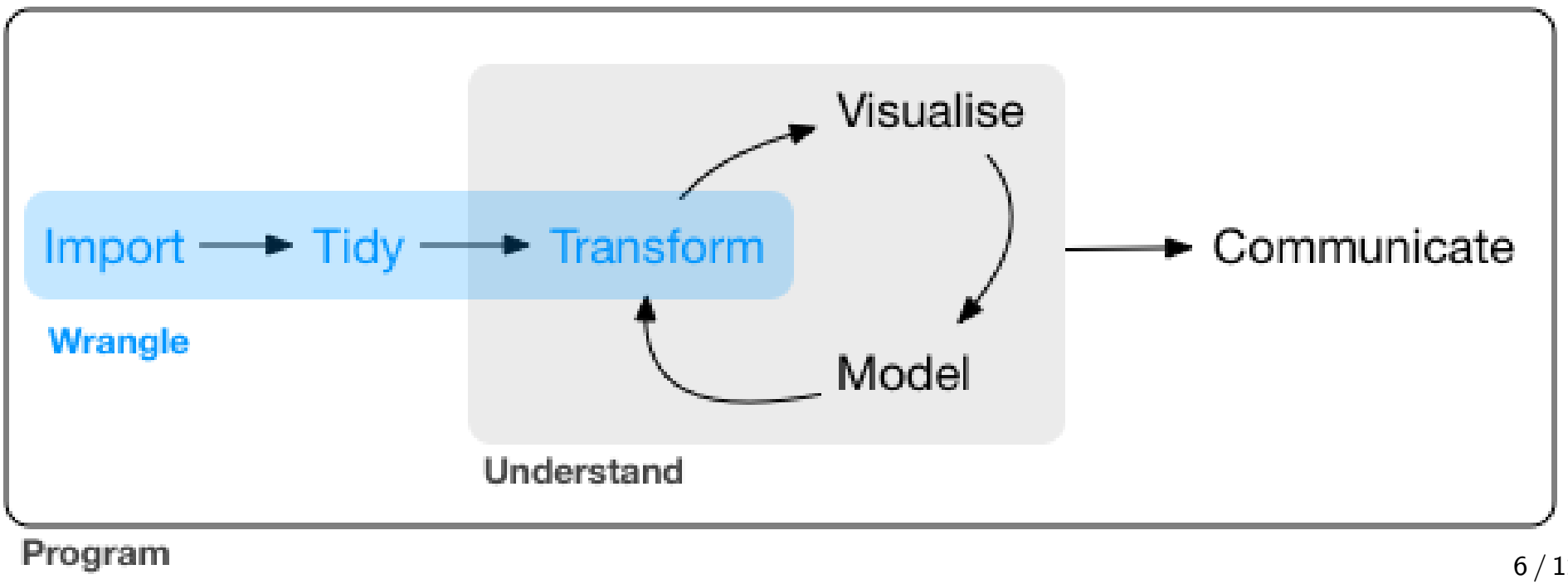
We have talked a lot about histograms and sampling distributions and have applied them to small data sets. Let's go through a larger running example and see some nicer visualizations.

Wrangling data

What is 'wrangling'?

From https://r4ds.had.co.nz/wrangle-intro.html:

> *In this part of the online book, you'll learn about data wrangling, the art of getting your data into R in a useful form for visualisation and modelling. Data wrangling is very important: without it you can't work with your own data! There are three main parts to data wrangling:*



Tibbles, data manipulation and introduction to graphics: a quick overview

A worksheet

Go to the "ToolsForEDA" file to walk through a detailed example. (posted in the same post as these notes!)

License

This document is created for ITMD/ITMS/STAT 514, Spring 2021, at Illinois Tech.

Appendix: programming

About pipes

https://r4ds.had.co.nz/pipes.html
https://stackoverflow.com/questions/24536154/what-does-mean-in-r

Functions, loops, etc.

Hands-on practice: go to Worksheet 6