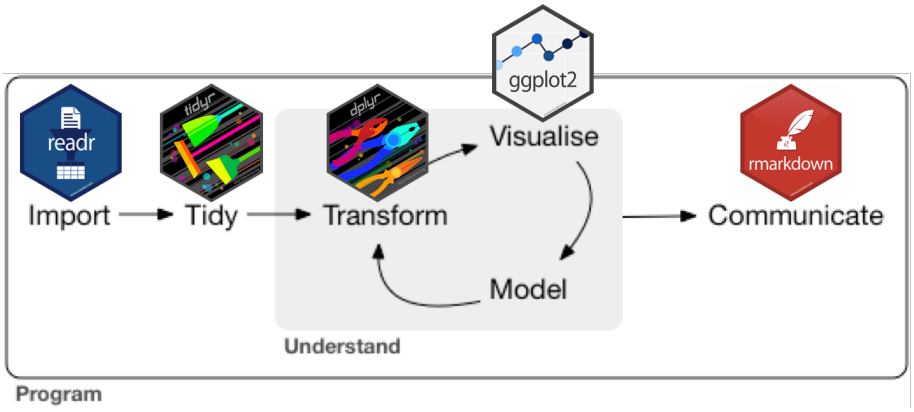
Unit 1 Review

September 13, 2022

Unit 1: R for data mining	**Lecture 1:** Intro to modern data mining
Unit 2: Prediction fundamentals	**Lecture 2:** Data visualization
Unit 3: Regression-based methods	**Lecture 3:** Data transformation
Unit 4: Tree-based methods	**Lecture 4:** Data wrangling
Unit 5: Deep learning	**Lecture 5:** Unit review and quiz in class

1 Introduction

In this lecture, we will review Unit 1, including data wrangling, manipulation, and visualization.



As usual, let's load the `tidyverse`:

```
library(tidyverse)
```

We will perform a data analysis to study trends in tuberculosis cases across the world over time. The two relevant datasets are:

- `who.tsv`: information about the number of tuberculosis (TB) cases from countries around the world each year during the period 1980-2013 (source: 2014 World Health Organization Global Tuberculosis Report)
- `population.csv`: the population of each country across time (source: The World Bank)

2 Import

First take a peek at the contents of `who.tsv` and `population.csv` by clicking on these files. Then, import the data into tibbles called `who` and `population`. How many rows and columns does each tibble have?

Look at the summary of the variable types printed by `readr` when importing `population.csv`. Does anything seem strange here? Fix the issue and store the result in a tibble called `population2`.

3 Tidy

3.1 `who`

Here are descriptions of the columns in `who`:

- `country`: the country name
- `iso2`: the two-digit country code
- `iso3`: the three-digit country code
- `year`: year
- variables like `new_ep_f014`: `ep` denotes the type of TB case (`rel` for relapse, `ep` for extrapulmonary, `sn` for smear negative, `sp` for smear positive); `f` denotes the sex (in this case female), `014` denotes the age group (in this case 0-14 years old).

3.1.1 Pivot

What are the variables in this data? What pivot operation would make the data more tidy? Apply this operation to `who`, and store the result in `who2`.

3.2 Separate

We need to separate out values like `new_ep_f014` into their constituent parts.

- As a first step, separate into values like `new`, `ep`, and `f014`. Remove the column containing `new`, as it contains no information. Store the result in `who3`.
- Separate values like `f014` into `f` and `014`. Store the result in `who_tidy`.

3.3 `population`

3.3.1 Pivot

What are the variables in this data? What pivot operation would make the data more tidy? Apply this operation to `population2`, and store the result in `population3`.

3.3.2 Cast

What type is the variable storing the population? What type should it be? Apply a casting operation to `population3` to remedy this issue, storing the result in `population_tidy`.

3.4 Join

What variable(s) should we use to join `who_tidy` and `population_tidy`?

- Apply a renaming operation to one of these two tibbles so that the variables used for the join have the same name.
- Join the two tibbles into a third called `tuberculosis`.

3.5 Clean up

Some of the variables in `tuberculosis` are unnecessary, and some of the values are `NA`. Remove the unnecessary variables and the rows containing `NA` values, storing the result in a tibble also called `tuberculosis`.

4 Manipulate

- How many total cases were there among men and women in the 21st century in America (`United States of America`)? Which sex had more cases?

- Create a new variable called `cases_per_100k`, which is the number of cases in a given year, sex, age group, and tuberculosis type per 100,000 people.
- Which country and which year had the most cases per 100k people (across all sexes, age groups, and tuberculosis types)? Which country and which year had the fewest cases per 100k people?

5 Visualize

- Plot the total number of cases per per 100k people as a function of year for the following three countries: China, India, and United States of America. Put the y axis on a log scale via `scale_y_log10()`. What patterns emerge?
- Compare the distributions of total cases per 100k people per country (summed over years, sexes, and tuberculosis types) across the different age groups. Put the y axis on a log scale via `scale_y_log10()`. What patterns emerge?
- Create a plot to assess whether, in the year 2000, the number of cases per 100k people in a country was related to the country's population. What do you conclude?