Lecture 2: Análisis EPH & Intro to Data Transformation

Economía Laboral

Junghanss, Juan Cruz

Universidad del CEMA

2nd Semester 2022

Contents

Today's lecture content:

- Introduction to LaTeX
- Introduction to Data Transformation
- Análisis de la EPH

Some announcements:

- The website for the course is (almost) ready. You can start using it: junghanss.github.io

Introduction to LaTeX

¿Qué es LaTeX y por qué usarlo?

- Sistema de composición de texto y lenguaje Markup.
- Utilizado ampliamente hace muchos años en el mundo académico.
- Mayor flexibilidad que los procesadores de texto formateado (ej: Word).
- Les da experiencia para otros lenguajes markup como Markdown.

Toda la info, templates y links están subidos en nuestra web:
https://junghanss.github.io/intro-latex/

Introduction to LaTeX

Les recomiendo trabajar con Overleaf en un principio, aunque es bueno que sepan que pueden tener LaTeX y un compilador localmente descargado en la computadora.

https://overleaf.com/

Inicialmente será bueno que tengan a mano templates, cheatsheets, manuales y los tutoriales de Overleaf para ir viendo cómo se hacen las cosas.

Intro to Data Transformation: working with data

How would you describe data? And a database? ... By its size, variables, observations?

Is the data also **tidy**? (We'll see how to tidy data in the next lecture).

Working with Data

Data can be defined and described in several ways. Let's begin by the most basic concepts:

1. Format:
   - **Quantitative**, i.e. numeric (continuous and discrete)
   - **Qualitative**, i.e. categorical or non-numeric (ordinal or nominal)
2. Structure:
   - **Structured data**: data that was predefined and formatted to a set structure before being storaged. Example: dates, phone numbers, addresses, product names, etc.
   - **Unstructured data**: data stored in its native format and not processed until it is used. Example: emails, social media, websites, sensor data...
   - **Semistructured data**: unstructured data with metadata[1] that identifies certain characteristics. Example: LaTeX

[1] Metadata is data about data.

Working with Data: Data Structures

Although this is not a programming course, you should know some definitions before working with data. You'll gain understanding that will help you work on professional projects and real-life problems.

What is a **data structure**? Basically, a format of data *organization*, *management* and *storage*. You have two main categories:

- **Static**: structures that are static (don't change), i.e. fixed in size.
- **Dynamic**: structures that change (grow or shrink) during runtime.

Data Structures

Some examples of data structures:

1. **Static**:
   - **Arrays**: they hold items of the same data type and you define the size when you create them.
   - **DataFrames**: they organize data into a 2D table of rows and columns.
   - **Tibbles** (in R): **lazy** and **surly** dataframes (we'll learn about them later)
2. **Dynamic**:
   - **Nested lists** (listas conectadas): linked lists (double linked or circular). Example: spotify playlist
   - **Stacks** (pilas): last in first out. Example: Undo/redo in Excel or Word.
   - **Queues** (colas): first in first out. Example: printer queue.
   - **Trees** (arboles): bidimensional nonlinear structure.

sequential, nonlinear, random access is not possible

Data Structures

Some examples of data structures:

1. **Static**:
   - **Arrays**: they hold items of the same data type and you define the size when you create them.
   - **DataFrames**: they organize data into a 2D table of rows and columns.
   - **Tibbles** (in R): **lazy** and **surly** dataframes (we'll learn about them later)
2. **Dynamic**:
   - **Nested lists** (listas conectadas): linked lists (double linked or circular). Example: spotify playlist
   - **Stacks** (pilas): last in first out. Example: Undo/redo in Excel or Word.
   - **Queues** (colas): first in first out. Example: printer queue.
   - **Trees** (arboles): bidimensional nonlinear structure.

Data Structures

Further definitions that are of our interest:

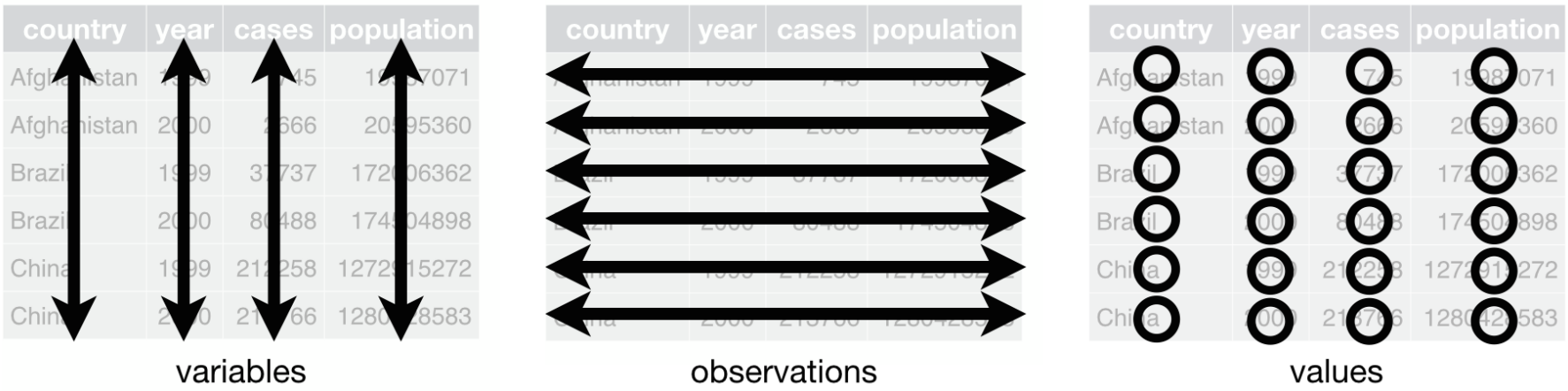
Rectangular data (structure): multivariate cross-sectional data in which each column is a variable (feature), and each row is an observation. These structures generally store data in a two-dimensional (2D) format (i.e., a grid containing rows and columns). For example: Excel, Google Sheets, etc.

It's like thinking about matrixes.

Tidy data

A dataset is **tidy** if it fulfills the following conditions:

1. Each variable must have its own column.
2. Each observation msut have its own row.
3. Each value must have its own cell.



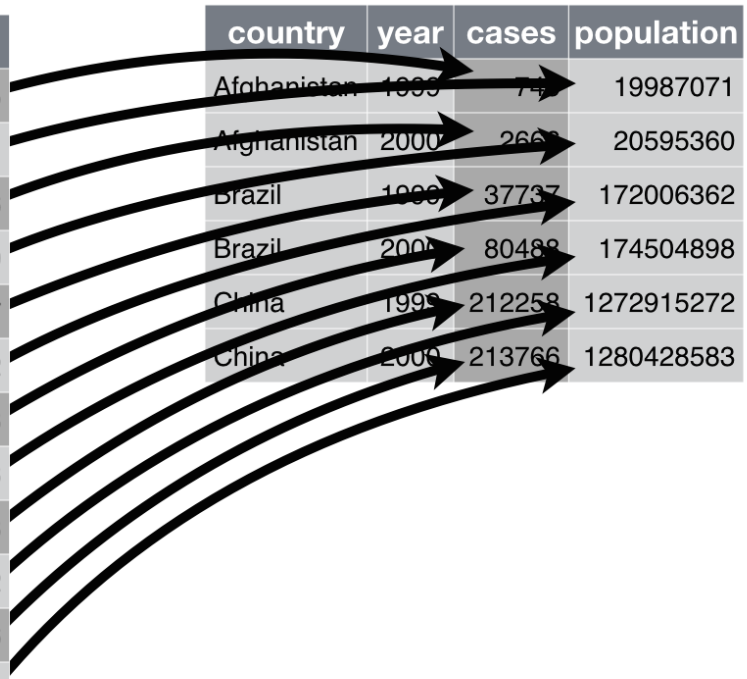
All the packages in the tidyverse environment are designed to work with tidy data.

Tidy data

Next class we'll learn how to tidy data, i.e. transform it, but in the meanwhile think about the following transformation. It will help you understand what is tidy data:

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2



Motivation

¿Por qué vemos estas definiciones sobre tipos, estructuras de datos, bases y demás?

El objetivo es que empiecen a pensar los problemas (desde un TP hasta desafíos profesionales) como programadores, además de como economistas.

Motivation: EPH

Hence, how can we describe the EPH?

1. Quantitative and qualitative data
2. Structured data
3. Rectangular data
4. Tidy data

Análisis de la EPH

La **Encuesta Permanente de Hogares**, como bien se indicó en la clase anterior, es desarrollada por INDEC y consta de una encuesta trimestral sobre una muestra de la población para obtener indicadores sociodemográficos, del mercado de trabajo, etc.

Hagamos un vistazo y luego vamos a RStudio a trabajar en el TP Nº1. Descarguemos las bases, el diseño de registro y el informe técnico[2]: https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos

[2] Inicio - Sociedad - Trabajo e ingresos - Mercado de trabajo

Análisis de la EPH

Características:

- Dataset "individual": datos sobre las personas.
- Dataset "hogar": datos sobre los hogares de los individuos.
- Ambos datasets están relacionados con unas keys: CODUSU y NRO_HOGAR.
- El dataset individual posee 177 variables y 49706 observaciones.
- En el "diseño de registro y estructura" tenemos info necesaria para identificar las variables, cómo están construidas, etc.
- En el informe técnico (Nº 115) sobre el mercado de trabajo, tenemos estadísticas descriptivas computadas a partir de los datos de esta EPH en cuestión.