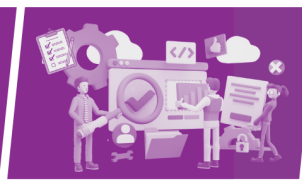
Fall 2023

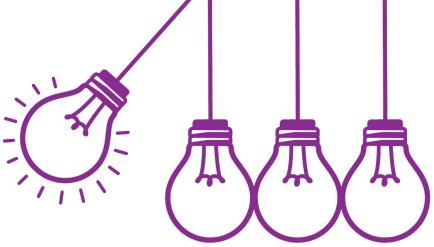
DIME Analytics

REPRODUCIBLE RESEARCH FUNDAMENTALS



THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

i2i DIME
TRANSFORM DEVELOPMENT



Tidying data - Hands on

Reproducible Research Fundamentals

September 26, 2023

Development Impact Evaluation (DIME)
The World Bank

- During the training, find all materials in our shared OneDrive: here

WORLD BANK GROUP

i2i DIME
TRANSFORM DEVELOPMENT

Importance of Tidying Secondary Data

Understanding Secondary Data

What is Secondary Data?

- Data collected by a party other than the user.
- Sources include government reports and big tech firms like META, Google, and Ookla.

Why Use Secondary Data?

- Leveraging existing resources for deeper insights.
- Can be more economical and quicker than primary data collection.

Quality Considerations

- **Reliability**: Scrutinize the source and its trustworthiness.
- **Authenticity**: Verify the data's authenticity and correctness.

The Importance of Tidying Secondary Data

Why Tidy Secondary Data?

- Ensures accurate analysis.
- Facilitates easier handling of data.

Appropriate Cleaning of Secondary Data

- **Spotting Errors Early**: Identifying discrepancies and anomalies at the outset.
- **Handling Missing Values**: Developing strategies for missing values.

Takeaway

- Tidy data supports accurate insights and informed decision-making.
- Adequate cleaning sets the stage for future research and reusability of the data.

Data

Data

This exercise utilizes two data sets.

- **Colombia's Connectivity**
  - File 1: `colombia_connectivity_wide.csv`
  - Source: Ookla and Humanitarian Data Exchange
- **Colombia/s infrastructure**
  - File 2: `colombia_infrastructure_lng.csv`
  - Source: Open Street Maps and Humanitarian Data Exchange

Exercise

Exercise 0

Through these hands-on lectures, you will work with two datasets, one from *Ookla* and one from *OpenStreetMap*. The objective of this exercise is for you to understand the data you will use.

Exercise 0: Familiarize with the Data

1. **Exploration:**
   - Visit the *Ookla.* and *OpenStreetMap* websites.
   - On Ookla: navigate to the table detailing the variables. Understand the metrics and how they represent connectivity.
   - On Open Street Maps. Review the different amenities. The amenities included in the dataset are ”school”, ”colleges”, ”hospitals”, ”clinics”, and ”universities”. But as you will see there are many more.

Exercise 0

Exercise 0: Familiarize with the Data

2. **Download and Preview in R:**
   - Read and preview both datasets.
   - Explore the datasets to understand the unit of observation, number of units, and the variables.
   - Note any missing values, special characters, the shape of the data, the differences (if there are) between the unit of observation.
3. **Reflect on next steps and possible applications:**
   - Based on your initial inspection, what potential issues can you foresee when tidying or cleaning the data?
   - How could you use this data in a project? How can having this type of data enrich our understanding of a region?

Exercise 1

- The folder you downloaded previously includes a template script you can use to write your solution for each exercise.

Exercise on Tidying Connectivity Data for Colombia

1. Read and preview the ‘colombia_connectivity_wide‘ dataset in R:
2. Remove duplicates. You can use *distinct*.
3. Open the help file for `pivot_longer` to understand its usage. Convert the ‘colombia_connectivity_wide‘ dataset into a long format using `pivot_longer`. Focus on columns related to metrics for different months (e.g., ‘avg_d_kbps_01‘, ‘avg_d_kbps_04‘, etc.).
4. Ensure that the resulting dataset has columns indicating the trimester, and the corresponding value.

Exercise 2: Tidying Infrastructure Data for Colombia's Municipalities

1. Load the necessary libraries in R:
2. Read and preview the 'colombia_infrastructure_long' dataset in R:
3. Explore the data
   - Which units of observation are included?
   - Which columns contain data of which units?
4. Open the help file for `pivot_wider` to understand its usage:
   - R: `library(tidyr)` and then `?pivot_wider`
5. Convert the 'colombia_infrastructure_long' dataset into a wide format using `pivot_wider`.

Challenge - How can tidyness help you?

Challenge

Analyze the data from exercises 1 and 2 to answer the following questions, comparing the ease of the process between using **tidy and untidy data**:

- From the ‘connectivity_long‘ and ‘connectivity_wide‘ datasets, which municipality (‘ADM2_ES‘) has the highest average download speed in the last trimester of 2020?
- Based on the restructured ‘colombia_infrastructure_wide‘ and original ‘colombia_infrastructure_long‘ datasets, which municipality (‘ADM2_ES‘) has the highest and second highest total count of schools, colleges, and universities combined?
- Why is more convenient using one or the other format in both cases?

The End