Fall 2023

DIME Analytics

Reproducible Research Fundamentals

September 25-29, 2023

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Fall 2023

DIME Analytics

Tidying Data Hands On

Benjamin Daniels

DIME Analytics

bdaniels@worldbank.org

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

Fall 2023

DIME Analytics

Tidying Data Hands On

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

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

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 (FB), Google, and Ookla.

Why Use Secondary Data?

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

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.

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

Data Exercise

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT

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

2. Download and preview data

- 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

Go to the materials folder and download the files indicated for each exercise:

The folder includes a template script you can use to write your solution.

Exercise on Tidying Connectivity Data for Colombia:

1. Open and preview the colombia_connectivity_wide dataset in Stata.
2. Remove duplicates.
3. Open the help file for [reshape] to understand its usage.
4. Convert the colombia_connectivity_wide dataset into a long format using [reshape].
   Focus on columns related to metrics for different months (e.g., ‘avg_d_kbps 01‘, ‘avg_d_kbps_04‘, etc.).
5. Ensure that the resulting dataset has columns indicating the trimester, and the corresponding value.

Exercise 2

Exercise on Tidying Infrastrcture Data for Colombia:

1.Open and preview the ‘colombia_infrastructure_long‘ dataset in Stata:

2.Explore the data:

- Which units of observation are included?
- Which columns contain data of which units?

1.Open the help file for [reshape] to understand its usage.

2.Convert the ‘colombia_infrastructure_long‘ dataset into a wide format using [reshape].

Challenge - How can tidy-ness help you?

Exercise on Tidying Infrastrcture Data for Colombia:

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?

Fall 2023

DIME Analytics

Tidying Data Hands On

Benjamin Daniels

DIME Analytics

bdaniels@worldbank.org

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

i2i DIME
TRANSFORM DEVELOPMENT