Fall 2023

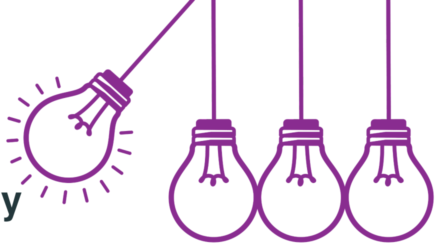
DIME Analytics

REPRODUCIBLE RESEARCH FUNDAMENTALS

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

i2i
DIME
TRANSFORM DEVELOPMENT

Data Construction - Secondary

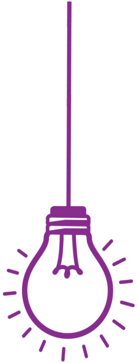


Reproducible Research Fundamentals

September 27, 2023

Development Impact Evaluation (DIME)
The World Bank

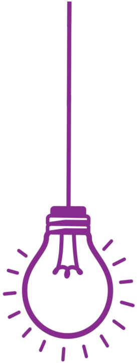
- During the training, find all materials in our shared OneDrive: here



Motivation

Motivation for Constructing with Secondary Data

- **Depth of Analysis:** Using secondary data can offer new perspectives and deeper insights into the primary data.
- **Validation:** Secondary data aids in validating and benchmarking results derived from primary data against established datasets.
- **Innovation:** Encourages creative and innovative approaches to data analysis.
- **Informed Decision-Making:** Facilitates more grounded strategies in policy and decision-making.
- **Collaborative Insight:** Allows for the combination of insights from different data sources at a low cost, improving decision-making.

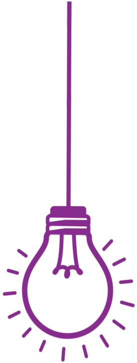


Data

Data

This exercise utilizes two data sets.

- **Colombia's Connectivity**
  - File 1: `colombia_connectivity_cleaned.csv`
  - Source: Ookla and Humanitarian Data Exchange
  - You created this in the cleaning, but is available in the data folder.
- **Colombia's Infrastructure**
  - File 2: `colombia_infrastructure_cleaned.csv`
  - Source: Open Street Maps and Humanitarian Data Exchange
  - You used a version of this dataframe in past exercises, but this has been cleaned to remove special characters so that both dataframes can be merged (you already cleaned the previous one).



Exercise

Exercise

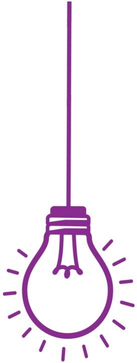
You will be working on a project aiming to analyze connectivity and infrastructure in Colombia. The goal is to produce detailed analyses across different administrative levels and types of infrastructure. The analysis will include:

1. View of the current state of connectivity in different regions.
2. A detailed breakdown of various types of infrastructure present in different regions.
3. Insights into the correlations between connectivity and infrastructure.
4. Quarterly analysis of connectivity performance metrics.

Exercise

To do this you will need to carry out the following taks. Here is a brief outline and more details on each task will follow later on the slides.

- Task 1: Plan construct outputs
- Task 2: Standardize units
- Task 3: Handle outliers
- Task 4: Create indicators
- Task 5: Create outputs data files



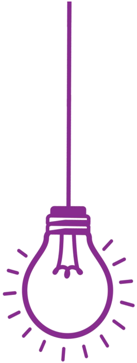
Task 1: Plan construct outputs

Plan construct outputs

Task 1: Plan construct outputs

- How many analysis data sets will you have to create?
- What are the unit of observations in each of them?

The solution to this task can be a short text, a few bullet points, a diagram etc.



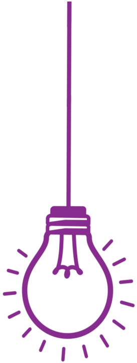
Task 2: Standardize units

Standardize

Our suggested best practice: *The conversion scalars may only be coded once. Meaning that if the conversions from, for example, KBps to Mbps would change, then only one place in the code should have to be updated.*

Task 2: Standardize Units

- Convert all speed measurements to Mbps.
- Ensure consistency in units across all datasets (if you create more than one).



Task 3: Handle outliers

Outliers - Task and discussion

Task 2: Deal with outliers

- Identify which connectivity variables have outliers
- Winsorize outliers for each connectivity type and trimester with more than 100 data points at the 99% level

Discuss:

- Should the original variable with outliers be overwritten?
- Why doesn’t it make sense to winsorize a variable with less than 100 data points at the 99% level?
- Compare distribution of `avg_d_kbps_04` with its winsorized version. Are there fewer observation that risk dominate the mean? Are there still such observations?
- Would it make sense to winsorize number of schools? What would be the issue with doing that?

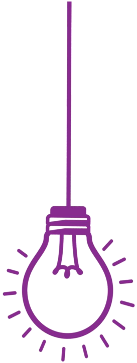
Outliers

Winsorization in Stata

Great package exists: `ssc install winsor`

Winsorization in R

There are packages with winsorization functions in R (see: `DescTools` and `HDRobust`), but to use `tidyverse` functions and to account for missing values, a custom function is needed. You can use the function on the template as a starting point.



Task 4: Create indicators

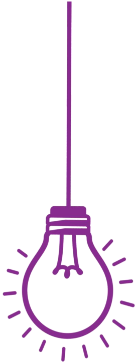
Combining variables to create indicators

Task 4: Create indicators

- Construct indicators that show average connectivity speeds (upload and download) per quarter and state and municipality.
- Create indicators for the number of the different types of infrastructure in each municipality and state.
- Develop quarterly change indicators of connectivity speeds (upload and download) by municipality.
- There are many tiles where the number of tests is limited. What would you do for these cases? Remember to document it.

Discuss:

- How can missing values be treated in the dataset to ensure accurate indicators?
- What potential insights can these indicators offer for decision-makers?

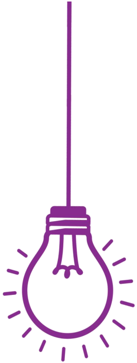


Task 5: Build checks in your code

Implementing Checks and Assertions in Your Code

Key Points to Remember:

- Check the number of rows and columns after each critical operation in your data manipulation script.
- Ensure unique keys maintain their uniqueness post operations like joins and merges.
- Use assertions to define and enforce expected value ranges, helping to maintain data integrity throughout the script.
- Leverage packages like ‘assertthat‘ to streamline the process of adding checks and assertions to your R scripts.



Task 6: Create outputs data files

Save analysis data sets

Remember this for next task:

Drop all variables apart from those needed in the analysis. It is easy to come back to this step if you drop too many. The fewer variables the lower the risk of confusing during the analysis stage.

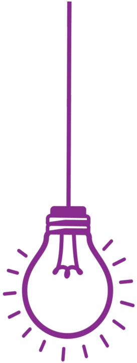
The only transformation an analysis script may do is to subset the data. For example, load the connectivity data, subset municipalities with school count, and then perform analysis.

Save analysis data set

- Create the minimal analysis data sets you think are needed in the analysis given the description above
- Make sure that the data sets are uniquely and fully identified, has labels and other documentation etc.

From earlier, the outcomes of interest are:

1. Average connectivity speeds per quarter and state and municipality.
2. Number of amenities per municipality and state.
3. Change in connectivity speeds.
4. Indicator showing both connectivity and infraestructure by municipality.



Thank you!