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

Data Construction Hands On

Benjamin Daniels, Roshni Khincha

DIME Analytics

bdaniels@worldbank.org

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

i2i DIME
TRANSFORM DEVELOPMENT

Fall 2023

DIME Analytics

Data Construction 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

Motivation for Secondary Data Construction

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 Exercise

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

i2i DIME
TRANSFORM DEVELOPMENT

Exercise

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 will be available in the data folder.

Colombia's Infrastructure

- File 2: colombia_infrastructure_wide.csv
- Source: Open Street Maps and Humanitarian Data Exchange • You used a version of this dataframe in past exercises, but this has beencleaned to remove special characters so that both dataframes can be merged(you already cleaned the previous one).

Exercise (continued)

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 (continued)

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

Exercise (continued)

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

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

Task 3: 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

Exercise (continued)

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?

Exercise (continued)

Task 4: Create indicators

- Construct indicators that show average connectivity speeds (upload and download) per trimester 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.
- Extra Challenge: Build a comprehensive indicator reflecting both connectivity and infrastructure data by municipality.

Exercise (continued)

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.

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 infrastructure by municipality.

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