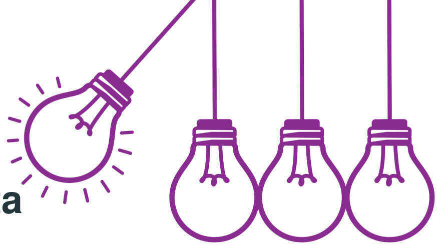
Fall 2023

DIME Analytics

REPRODUCIBLE RESEARCH FUNDAMENTALS

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

i2i
DIME
TRANSFORM DEVELOPMENT



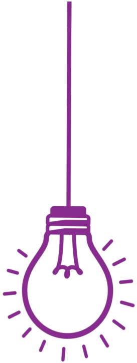
Data Analysis: Secondary data

Reproducible Research Fundamentals

September 28, 2023

Development Impact Evaluation (DIME)
The World Bank

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



Data analysis: objectives and data

Objective of the Course

In this course, you will learn to leverage secondary data to unearth answers to crucial research questions. The analytical phase is the bridge from raw data to discerning insights.

We will predominantly use **R** to explore various analytical techniques. We will use the data resulted from the construction exercise.

- **Municipality Database**:
  - `municipality_database.csv`
  - Source: Ookla, Open Street Maps and Humanitarian Data Exchange.
- **State Database**:
  - `state_database.csv`
  - Source: Ookla, Open Street Maps and Humanitarian Data Exchange.

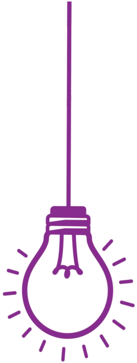
Tasks

The exercise will consist in the following tasks.

- **Task 1 Summary statistics**: Describing the central tendencies, dispersion, and shape of a dataset's distribution.
- **Task 2 Boxplots and histograms**: Visualizing the distribution of a dataset.
- **Task 3 Regression analysis**: Understanding the relationships between variables.
- **Task 4 Visual analysis using ggplot**: Crafting informative visualizations.
- **Task 5 Correlation analysis**: Finding relationships between different variables.

Other potential steps in data analysis

- **Time series analysis**: Analyzing data collected over a period.
- **Cluster analysis**: Grouping similar objects together.
- **Principal Component Analysis (PCA)**: Reducing data dimensionality while retaining information.



Statistics, dispersion and regression tables

Task 1: Summary Statistics

Objective
In this task, we aim to understand the central tendencies, dispersion, and shape of our dataset's distribution using summary statistics.

Steps

1. Load the necessary R packages: `stargazer` `ggplot2`. The ones proposed are already in your template file, but feel free to use others.
2. Calculate the summary statistics.
3. Export the summary statistics table to a LaTeX or HTML file using the `stargazer` or `gt` package.

Task 1: Suggestions

In R, creating summary statistics tables can be a bit more manual compared to other statistical software such as Stata. However, this allows for high customization to suit your specific needs. In your template, one method is proposed, but we encourage you to explore and possibly find more convenient or suitable options for your task.

Suggested packages and functions

- `dplyr` and `tidyr` for data manipulation and transformation.
- `gt` package for creating beautiful and highly customizable tables.
- `stargazer` or `kable` from the `knitr` package, which is another popular choice for creating tables.

Task 2: Boxplots and Histograms

Objective
Visualize the distribution of different variables in the dataset using boxplots and histograms.

Steps

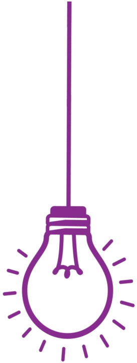
1. Use ggplot2 package to create boxplots and histograms in R.
2. Identify outliers and understand the distribution of your data through boxplots.
3. Get a sense of the central tendency, variability, and the shape of the distribution of your data through histograms.
4. Save the most relevant figures, if there are any using `ggsave()`.

Task 3: Regression tables report

Objective
Understand the relationships between variables through regression analysis.

Steps

1. **Selecting Variables:** Select dependent (download and upload speeds) and independent variables (e.g., number of schools).Think why this could have endogeneity problems, and then for the purpose of this excercise ignore it.
2. **Building the Regression Model:** Utilize the `lm()` function in R to build your regression model using the municipality database.
3. **Analyzing the Model:** Use the `summary()` function in R to get detailed information on your model.
4. **Add clusters:** Add clustered standard errors at state level.
5. **Save it:** Save all your models using stargazer.



Visual analysis

Task 4: Visual Analysis using ggplot

Objective Using graphs can help bring insights and patterns into a clearer focus. They increase the communicative power of your data, translating numbers into visuals that can be more intuitively understood.

Steps

1. **Create scatter plots:** Use ggplot to analyse the relationship on the previous task. Use `geom_smooth(method = "lm"`
2. **Create bar:** Find which states suffered more in terms of connectivity from 2020Q1 to Q4 and plot it.
3. **Interpretation:** Do any other graph that could help you to deduce patterns, trends, and insights and save your graphs.

Tips, Tricks, and Suggestions

Enriching your Visualizations:

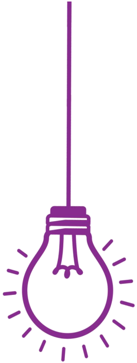
- **Layering:** Incorporate different layers (like `geom_point`, `geom_line`).
- **Faceting:** Utilize facets to create a matrix of plots. You can use `facet_wrap`
- **Themes:** Apply different themes to tailor the aesthetics of your plot.
- **Color:** Consider adding color, size, and shape aesthetics to enrich the information shown.

Graph Types and their Utilities:

- **Scatter Plots:** Visualizing relationships between two continuous variables.
- **Bar Charts:** Distribution of a categorical variable.
- **Histograms:** Distribution of a single continuous variable.
- **Box Plots:** Snapshot of the data's central tendency and spread.

Task 5: Correlation Analysis

- **Objective**: Understand the relationships between different variables using correlation analysis. Here, we will focus on analyzing the relationship between social infrastructure and average connectivity speed.
- **Tasks**:
  1. Filter the necessary data from the municipality database for correlation testing.
  2. Perform a correlation test using the `cor.test()` function in R.
  3. Create a correlation matrix using selected variables from the state database.
  4. Visualize the correlation matrix using the `corrplot()` function to represent correlations graphically.



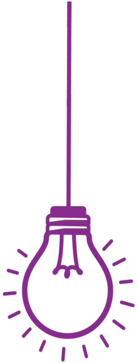
Resources

DIME resources

- Development Research in Practice - https://worldbank.github.io/dime-data-handbook/analysis.html
- R Econ Visual Library - https://worldbank.github.io/r-econ-visual-library
- Stata Visual Library - https://worldbank.github.io/Stata-IE-Visual-Library
- DIME LaTeX training - https://github.com/worldbank/DIME-LaTeX-Templates
- Checklist: Reviewing graphs - https://dimewiki.worldbank.org/Checklist:_Reviewing_Graphs
- Checklist: Reviewing tables - https://dimewiki.worldbank.org/Checklist:_Submit_Table

External resources

- Grids of Numbers (Butterick): https://practicaltypography.com/grids-of-numbers.html
- Common Issues in Exhibits (JCE): https://www.jclinepi.com/content/checklist_for_tables_and_figures
- Data Visualization Checklist (Evergreen): https://stephanieevergreen.com/data-visualization-checklist
- Accessible Data Visualization (Organ): https://towardsdatascience.com/an-incomplete-guide-to-accessible-data-visualization-33f15bfcc400



Thank you!