**DAPP Final project – Data cleaning process**

**Topic**: Relationship between conditional cash transfer programs and education in rural areas of Latin America.

Although the literature widely supports conditional cash transfer as an effective policy strategy to combat poverty in developing countries, we want to see if these results are differential among urban and rural communities. Our question comes from the concern that rural communities often do not have the infrastructure that allow potential participants to meet the conditions of transfer programs. For instance, if a family lives in a community that does not have a school, a program that mandates school enrollment will exclude them from its benefits. With this project, we hope to gain a better understanding of the impact of conditional cash transfers specifically on rural communities.

We plan to use datasets compiled by the [Center for Distributive, Labor, and Social Studie](https://www.cedlas.econo.unlp.edu.ar/wp/en/)s, from the National University of La Plata, which are publicly available in their website for download. With this data, we hope to create interactive scatter plots and maps, displayed in a well-organized dashboard.

**Folder structure in GitHub:**

Structure under the folder “data”:

**data/raw:**

* 2024\_Act1\_enrollment\_LAC.xlsx - [School attendance](https://www.cedlas.econo.unlp.edu.ar/wp/wp-content/uploads/2024_Act1_enrollment_LAC.xlsx)
* 2024\_Act1\_literacy\_LAC.xlsx –
* 2024\_Act1\_years\_edu\_LAC.xlsx - [Years of education](https://www.cedlas.econo.unlp.edu.ar/wp/wp-content/uploads/2024_Act1_years_edu_LAC.xlsx)
* 2024\_Act1\_years\_edu\_LAC.xlsx - [Infrastructure](https://www.cedlas.econo.unlp.edu.ar/wp/wp-content/uploads/2024_Act1_infrastructure_LAC.xlsx)
* 2024\_Act1\_housing\_LAC.xlsx - Housing

**data/cleaned:**

* all\_education.csv
* all\_infrastructure\_housing.csv
* cct\_countries.csv

**Final datasets** (Already in github)**:**

* all\_infrastructure\_housing.csv
* all\_education.csv

\* Please, drop all dataset outside of the folder “data”.

1. **Cleaning process:**

Each country has data from different data, depending on the source of information. For example, for the file “2024\_Act1\_enrollment\_LAC.xlsx”, Argentina has information from 1980 until 2022, but the source it is not the same since the beginning. See the image:

Imagen que contiene Tabla

Descripción generada automáticamente

Then, in the cleaning process for each country it is necessary to choose what source use. For all countries with CCT, we are keeping information for years before the implementation, beside the source of information are different. For each country this process was followed:

* Argentina:
  + We are using the Continuous Permanent Household Survey (EPHC: Encuesta Permanente de Hogares) second semester, which start in 2003 in the dataset.
* Bolivia:
  + We are keeping from 2011 onwards because there is not data for 2010.
* Brazil:

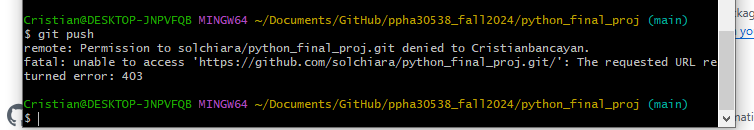
We are considering data from:

* + 1995 to 2003 under the New PNAD. They ask enrollment to all individuals.
  + 2004 to 2011 under the New PNAD With Rural North.
  + 2012 onwards under the PNADC. They capture enrollment from 5 years old.
* Chile
  + We are considering all available information. Not information for all years, it is every two years.
* Colombia
  + There is **no** information before the implementation of the CCT.
  + Data from 2001 to 2005 for ECH-National. In 2008, a new dataset named GEIH-National is available. Gap of 2 years (2006 and 2007).
  + There is a new methodology for the GEIH since 2021. 2021 and 2022 with new methodology is not considered. It would be just including one more year (2022).
* Costa Rica
  + Enrollment from 3-5 years old were asked from 2006. All the other information come from same source.
* Dominican Rep.
  + We set an arbitrary cutoff 2000. When a change in the methodology happened.
* Ecuador
  + Data from 2003, when ENEMDU started.
* El Salvador
  + We set an arbitrary cutoff 2000.
* Guatemala
  + We remove 2002, 2003 and 2004 because they are from another survey.
* Honduras
  + We keep information from 2001 since INE started.
* Mexico
  + Keep all the years. One source. Note: In 2016 there was a new methodology.
* Nicaragua
  + Keep all the years. One source.
* Panama
  + Keep all the years. One source.
* Paraguay
  + Keep all the years. One source.
* Peru
  + Keep all the years.
  + Note: ENAHO had many changes between 2000 and 2003. Maybe is better to use from 2003 onwards, but the CCT was implemented in 2005.
* Uruguay
  + Keep all the years. One source.
* Venezuela
  + Keep all the years. One source.

The structure of the sources per each country is similar in all the files from SEDLAS. Then, we applied the same code for all the files for Education but made some changes for files of Infrastructure (i.e., there is no information for infrastructure for Argentina).

**Pending:**

Made commit throughout git bash, be it appear this message:



Additional notes:

Tabla

Descripción generada automáticamente

Análisis de lenguaje:

* Discursos políticos alrededor de PTM
* Noticias en los países cuando se implementation los CCT
* Análisis de normas al momento que se implementación de los CCT