Final Project Proposal: Intro to Data Science, Spring 2024

Asad Azhar, Juan Menendez, Priscila Stisman, Talia Stringfellow

2024-04-15

Final project Proposal

Instructions: The proposal should be 2-3 paragraphs and include the names on the team, a brief description of the motivating question, all data sources that will be used, a description of the approach, and the two or three of the biggest anticipated technical hurdles in the project.

Motivating Question

Does self reported race/skin tone affect attained education level differently in different counties of South America?

Data Sources

1. Americas Barometer - LAPOP/ Vanderbilt University

Link to dataset: https://www.vanderbilt.edu/lapop/studies-country.php

The Americas Barometer is a periodic survey study of 34 countries in the Western Hemisphere, with stratified nationally representative samples drawn in each country, a common questionnaire core, and country-specific modules. It is the only scientifically rigorous comparative survey of democratic values and behaviors that covers all independent countries in North, Central, and South America, as well as a significant number of countries in the Caribbean. The Americas Barometer measures attitudes, evaluations, experiences, and behavior in the Americas using national probability samples of voting-age adults. This dataset contains our variable of interest: skin color. They ask the interviewed to self-select the color that best matches their skin color from a palette of 10 possible colors.

How data was created: nationally representative surveys across 34 countries in the American continent, from 2004 until now.

Data dictionary: https://www.vanderbilt.edu/lapop/ab2023/AB2023-Core-Questionnaire-V9.3-Eng-230228-W.pdf

Unit of analysis: person (universe: voting age adults) Data dimensions: for 2023, the dataset is composed by 43,074 cases, from 26 countries, covering a core questionnaire of around 200 variables (2023 core questionnaire). There is also data avaliable for: 2021, 2018/19, 2016/17, 2014, 2012, 2010, 2008, 2006, 2004.

Limitations: sampling errors / survey methods / some relevant political questions not included in the core questionnaire

Additional resource to interact with the data ('kid friendly'): https://www.vanderbilt.edu/lapop/interactive-data.php

2. World Bank's Education Statistics

Link to dataset: https://databank.worldbank.org/source/education-statistics-%5e-all-indicators/preview/on

This data set includes data for both spending, and educational outcomes for different countries over the years.

How data was created: This data is collected by the World Bank, through either National Administrations or surveys sponsored by the Bank. The MetaData includes details of all countries.

Data dictionary: metadata is available for download via Excel file.

Unit of analysis: country level data

Data dimensions: yearly data available - will use multi-year data based on what is appropiate in the context of the previous dataset. Limitations: It will have missing values for various countries depending on the year.

3.South American countries Geographic Data

Link to dataset: https://tapiquen-sig.jimdo.com/english-version/free-downloads/south-america/

How data are created: shapefile from South America in WGS84 Datum. Data dictionary: Columns are well defined because this is geospatial data.

Unit of analysis: Countries level (South America)

Dimensions of the data: polygon, geographic

Missing in the data: Nada

Limitations: Nada

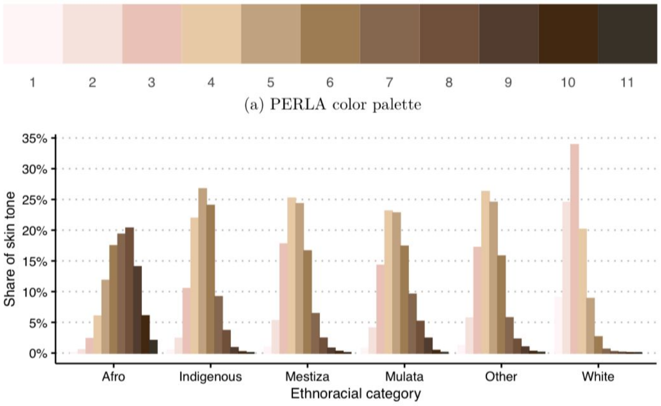
Sourcing:

“Shape downloaded from http://www.efrainmaps.es. Carlos Efraín Porto Tapiquén. Geografía, SIG y Cartografía Digital. Valencia, Spain, 2020.”

Description of the approach

Our approach will be rooted in econometric data analytics. We are interested in using the self-reported skin tone dataset to examine how educational outcomes are affected by racism differently across neighboring South American countries.

In order to do this we are first going to model the relationship between highest level of education achieved and skin tone within each country uniquely. What is interesting is that (as you can see in the figure below) there are two categorizations for skin tone: both ethnoracial category (6 options) as well as a purely skin tone based groupings (11 options). We will test both variables to see which is more strongly related to educational outcomes. We are interested in potentially using PCA methodologies to combine the effects. While doing state level data, we would control for urban/rural, gender, age and other factors that impact educational outcomes.



(a) PERLA color palette

(b) Ethnoracial categories and skin tone

“‘

After this initial modeling, we are interested in going to the national level to compare between neighboring countries. Here, we would control national level metrics (such as GDP, population age and size) to control for variation between countries. Our ideal outcome would be a map of South America to see how race affects educational outcome relative to neighboring countries

This project will use data visualization, geospatial analysis and potentially supervised machine learning.

Expected Technical Hurdles

We are concerned about reconciling the units of analysis between the datasets. Moreover, we are wondering how to integrate supervised machine learning, this research is mostly based on regression analysis rather than prediction. We are also expecting that the standardization of the data from country to country (that we will be required to do when producing our final deliverable of a comparative heat map).