Delving into the Determinants of Making it to University in South America

Data Science for Public Policy

Asad Azhar, Juan Menendez, Priscila Stisman, Talia Stringfellow

Summary

Georgetown University

1. Background and Literature Review

2. Data Sources

Americas Barometer - LAPOP/ Vanderbilt University

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 all our variables of interest

The AmericasBarometer by the LAPOP Lab, www.vanderbilt.edu/lapop

Rgeoboundaries - Runfola et al, 2020

For the geospatial analysis, we use data from rgeoboundaries. It contains political administrative boundaries for the globe as they were by 2020. Since South America hasn not experienced boundaries changes in 21st century, we use the dataset as it is, just adjusting the scope to the subcontinent.

Runfola D, Anderson A, Baier H, Crittenden M, Dowker E, Fuhrig S, et al. (2020) geoBoundaries: A global database of political administrative boundaries. PLoS ONE 15(4): e0231866. https://doi.org/10.1371/journal.pone.0231866

3. Data Wrangling

4. Exploratory Analisis: higher education in South America

5. Data Analysis: using unsupervised machine learning to predict attending to college in Argentina

6. Discussion of Results

Feedback

- need municipal level polygons/ geographic data for south america
- what kind of cleaning can we do for multicoliniarity?
- what kind of methodology can we use with time/year data?
- find another dataset that complements the data that you found
- where people live and how far is the neighborhood to each school
- start with bigger picture custom functions in purr and pass each name of country + look at relative feature importance

Data visualization

n 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.

geospatial analysis

This requires going in depth on each method you choose to demonstrate in your project (aka if you're going to choose geospatial analysis as one of your methods, I'd want to see you go beyond mapping).

supervised machine learning

- don't use model for education but rather look into how the model is using race/ethnicity 3 models: y = education data frame with ethnicity data frame with skin tone data frame without either variable and see how it affects precision and accuracy?

- **unsupervised machine learning**

I'd suggest you think about how to use the methods from the course to explore the topic.

You could use supervised machine learning to predict educational outcomes using skin tone and other variables. You could then examine the model feature importance to assess the predictive value of skin tone.

However, this is not the same as assessing the effect of skin tone on educational outcomes (that's the difference between a causal inference framework and a prediction framework). I could also see the use of cluster analysis to create groupings based on educational outcomes, skin tone, and other variables across countries. For the between-country analysis, I think you could do some extensive data visualization to explore those outcomes. You could also do a deep-dive on a couple countries of interest to use geospatial analysis to look at subnational patterns. You could also bring in different types of country specific data (e.g. the locations of schools) to enable different types of geospatial analysis.