DS4A COLOMBIA

Final Project Report

Gender Inequality in Colombia
Bogotá D.C., Colombia
Diciembre 10, 2019

Presented by:

Team 32
Angelica Mora
Andrés Argúmero
Carlos Sanmartín
Daniel Perico
Emiro Chica
José Campo

CONTENT

FIGURES

1 INTRODUCTION

1.1 Context

The Colombian state through the Dane (National Department of Statistics) conducted a population and housing census in 2018. The census consisted of counting and characterizing people residing in Colombia, as well as their homes and homes throughout the national territory. This allowed obtaining first-hand data on the number of inhabitants, their distribution in the territory and their living conditions.

> *"In this context, the information generated by the National Census of Population and Housing 2018 on characteristics of the population such as sex, age, ethnicity, cultural level, economic situation; and their respective living conditions, such as the conformation of households, head of household, types of housing, and access to public services, becomes essential information for the development of the country, and constitutes the main input to determine the evolution of demographic variables. For example, the size of the households, the index of aging, the index of youth, the migratory phenomena within the country and, from and to the outside, to name a few."*[1]

The census is not only a snapshot of the country at a given time, but also allows statistical information to be generated so that different public and private organizations in the country plan and make decisions on public policy, economic development, social welfare, employment, housing, Health, migration, among others.

The open data policy of the Colombian state makes the final data of the 2018 census available to the public.

National Census of Colombia 2018:

https://www.dane.gov.co/

Official Data from population in Colombia:

https://www.datos.gov.co/

1.2 Problem to solve

Based on a preliminary analysis of the 2018 census information, we observed a trend in gender inequality, related to the variables identified as indicators of well-being; the gap between men and women changed between the municipalities of each department, so we wrote our hypothesis as follows:

Which factors contribute to gender inequality in Colombia?

1.3 Project Scoping

Analyze the factors that influence the gender-based inequality based on the official data from government in Colombia.

Version 1: Analyze gender inequality respect to education and employment status factors in the Colombian population.

Version 2: Additionally to Version 1, analyze gender inequality respect to health services in the Colombian population

Version 3: Additionally to Version 2 analyze gender inequality factors by geographic zones in Colombia

1.4 Justification

Gender inequality exists in all countries of the world, in all social groups and is relevant as perpetuators of poverty. A girl who is born in a poor home and is forced to marry at an early age, for example, is more likely to drop out of school, give birth too young, suffer complications during childbirth and suffer domestic violence, than a girl of a higher income home.

On the other hand, the unequal remuneration between men and women is largely due to the disproportionate burden of unpaid domestic work faced by women, especially during their reproductive years.

In our country we cannot talk about development if we fail to reduce the gender gap that keeps women; this is directly affected by the stagnation of other factors such as poverty, hunger, health, education and employment, among others.

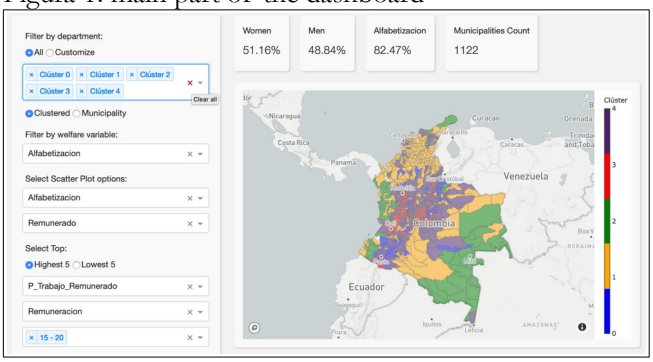
2 APPLICATION OVERVIEW

The front end of our application consists of a dashboard with graphic visualization of the 2018 census data.

The main part is the map of Colombia with the municipalities, colored according to the conventions of the clusters, which are a rank of the municipalities according to the values of the variables associated with well-being. The ranks shown on the map are fully customized, and the user can choose all or some of them and the map interactively shows only the municipalities of the chosen cluster. When you mouse over the municipalities, you can obtain its name as well as the population and the rank to which it belongs. Rank

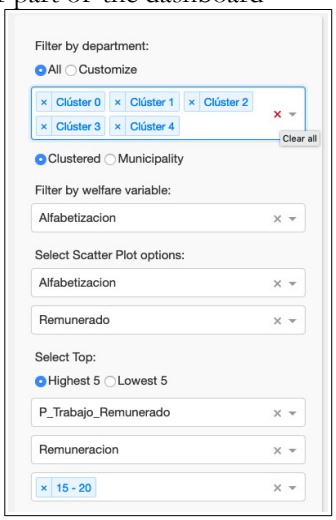
2 has the lowest values of these variables, while rank 3 has the highest values. Ranks 0, 1 and 4 have intermediate values.

Figura 1: main part of the dashboard



As a complement, there are five plots (barplots, scatter plot and line plot) that present the information of the different variables chosen, grouped by rank or showing the differences by sex.

Figura 2: filter part of the dashboard



The importance of the plots presented is that the user can propose different hypotheses about the inequality between men and women, according to the chosen variables, not only throughout the country, but between ranks of municipalities and be able to verify it in real time.

3 TECHNICAL EXPOSITION

3.1 Interactive Front-end:

3.1.1 Technologies to set up the front-end

The front-end is etirely based on Dash by Plotly. Dash is ideal for building data visualization apps with highly custom user interfaces in pure Python. Since Dash apps are viewed in the web browser, Dash is inherently cross-platform and mobile ready. Additionally some HTML and CSS tune up was necessary.

3.1.2 Front-end visualizations

The Dashboard of our application is made up of several interactive visualizations that allow us to graphically understand the socioeconomic characteristics of the Colombian population.

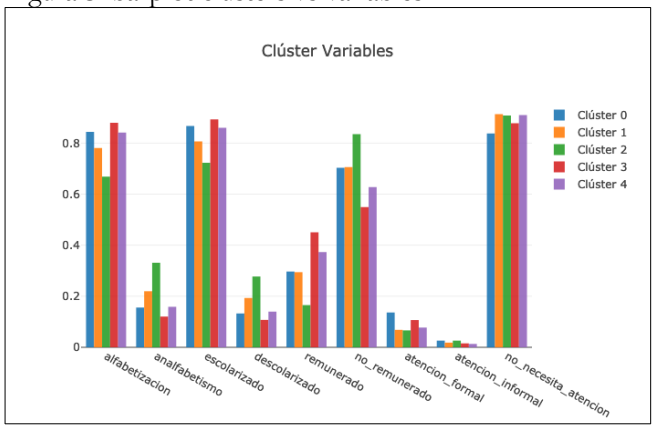
Map of Colombia containing the shape of the municipalities, showing with colors to which of the five clusters it belongs; We wanted to group the municipalities that show similar characteristics in their socioeconomic variables, in a few categories, identifying the two categories that will show an important gap in terms of inequality.

The dash also includes five types of graphs that complement the map information; These are interactive and allow the user to choose different socioeconomic variables, so that it is possible to investigate, establish new relationships between them and draw conclusions from their results.

3.1.2.1 Plot 1 bar plot

It shows a barplot by interactively chosen clusters of the dashboard, with the percentage of inhabitants in the four socioeconomic variables identified as indicators of well-being, these are literacy, schooling, health care and paid work.

Figura 3: barplot clusters vs variables

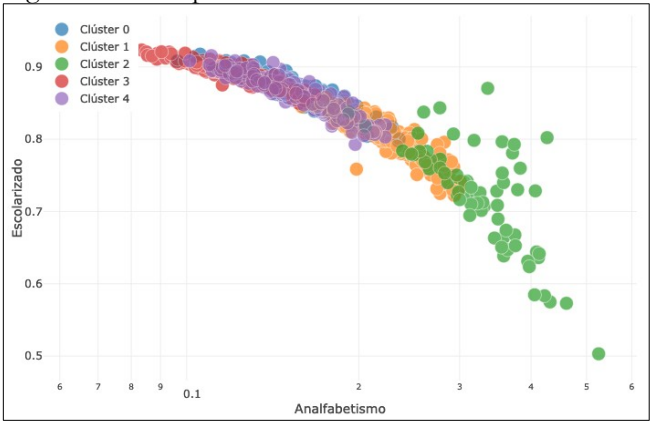


3.1.2.2 Plot 2 scatter plot

It shows a scatterplot by clusters, indicating the relationship between two socio-economic variables chosen interactively from the dashboard, identified as

indicators of well-being, these are literacy, schooling, health care and paid work.

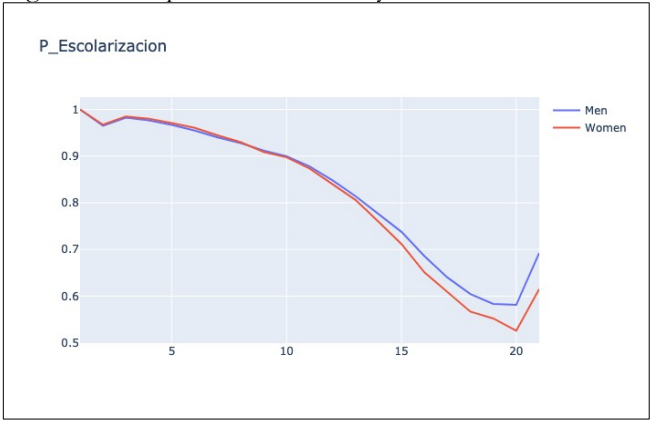
Figura 4: scatter plot – covariate ratio



3.1.2.3 Plot 3 line plot

It shows a lineplot, indicating the difference by sex of one of the socio-economic variables chosen interactively from the dashboard, of those identified as indicators of well-being, these are literacy, schooling, health care and paid work, throughout the different age groups.

Figura 5: line plot – difference by sex



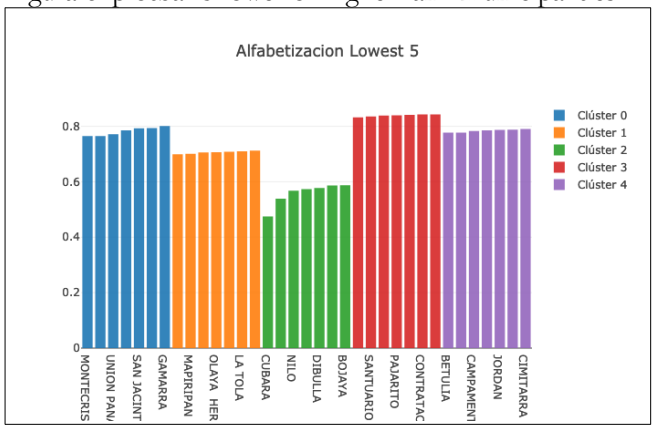
3.1.2.4 Plot 4 bar plot

It shows a barplot of the five clusters, indicating the five municipalities with the highest or lowest indicators (according to what was chosen in the dashboard) of the socioeconomic variable chosen interactively from the dashboard, of those identified as indicators of well-being, these are literacy, schooling, health care and paid work.

3.1.2.5 Gráfica 5 bar plot

It shows a barplot that indicates the difference by sex, grouped by clusters of the categories of the socioeconomic variable chosen interactively from the dashboard, from those identified as indicators of well-being, these are literacy, schooling, health care and paid work.

Figura 6: plotbar 5 lower or higher rank municipalities



3.1.3 Information flow between AWS-hosted components and Dash instance

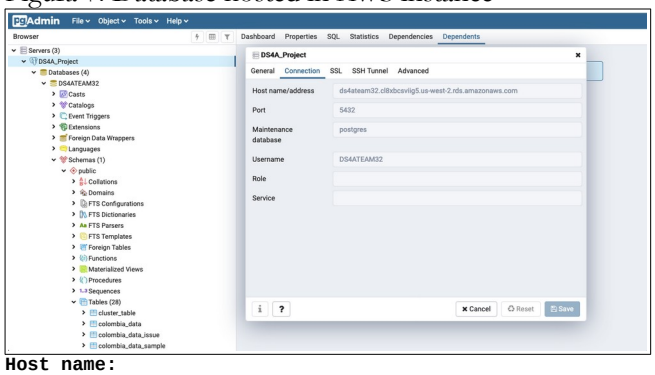
The way in which we connect the AWS EC2 instance with our Dash components was creating our own environment for the project in the EC2 instance and there we installed all the necessary dependencies so that our source code could run without any problem. Subsequently, all the files were transferred to the instance and started running in one of the ports together with the host that AWS gave us.

3.2 AWS-hosted Database:

3.2.1 Type of database

PostgreSQL 11.5 on x86_64-pc-linux-gnu, compiled by gcc (GCC) 4.8.3 20140911 (Red Hat 4.8.3-9), 64-bit AWS instance.

Figura 7: Database hosted in AWS instance



Host name:

ds4ateam32.cl8xbcsviig5.us-west-2.rds.amazonaws.com

Ddatabase name: DS4ATEAM32

3.2.2 Main datasets included

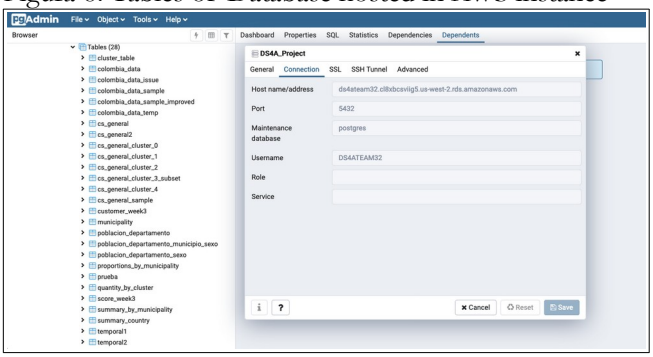
Raw table with all the 2018 census data, of *Personas* information. The *Personas* table contains the information of the records of people residing in private households with the characteristics corresponding to the census.

3.2.3 Main data tables set up

Summarized table with all the required metrics in absolute values and percentages.

```
Main data table: POBLACION_DEPARTAMENTO_SEXO
```

Figura 8: Tables of Database hosted in AWS instance



3.2.4 Tables Data Base Design

Design was based to place the most amount of workload in the database engine.

3.3 AWS-hosted Data Analysis & Computation:

3.3.1 Computation tool

A virtual machine instance hosted by EC2 AWS; x86_64-pc-linux-gnu, compiled by gcc (GCC) 4.8.3 20140911 (Red Hat 4.8.3-9), 64-bit.

This component is responsible for hosting the python scripts that extract the information from the main tables of the database hosted on the AWS instance, performs the calculations, runs models and transfers them to the dash, whose python script is also hosted in this instance.

3.3.2 Computacional package

The following are the main python scripts that connect the information between the database hosted on the AWS instance and the EC2 instance and allow dash execution:

3.3.2.1 mapa_municipios_live.py

This script hosted on the EC2 instance is responsible for running the dashboard and communicating with the tables hosted on the AWS Database instance.

3.3.2.2 kmeans_proyecto_01.ipynb

This script executes the K-means clustering using K-mean Sklearn library.

3.3.2.3 Query.ipynb

This script runs SQL queries from the main database hosted in the AWS instance to summarize tables that power the dash.

3.3.3 Data wrangling & cleaning process

Most of the wrangling and cleaning was done directly in SQL and consisted in reviewing the data, comparing it to the master information in the PDF provided by the DANE and setting up the proper data types for each one of the columns.

Proper care was taken to handle the invalid values to ensure that the proportions were properly maintained.

We summarized and pivoted the table, got subtotals by concepts and did proportions based on those. This was to reduce the original dataset of 40+ million rows into a table of 200 rows (at cluster level).

No advanced feature engineering was performed as most of the datas was categorical. Logical thinking to group data through visual means and correlation was done.

3.3.4 Models

During the analysis of the data we realized that there were differences in the behavior of the welfare variables between municipalities and we wanted to check if it was possible to group them by similar characteristics. We decided to use the K-means method because it matched the type of data we had. Grouping allowed us to analyze and present the information in a clearer way and conclude that in Colombia we have high and low ranking municipalities, with a concentration of the first ones close to the big cities and a dispersion in the borders of the latter.

Figura 9: Elbow method for determining K

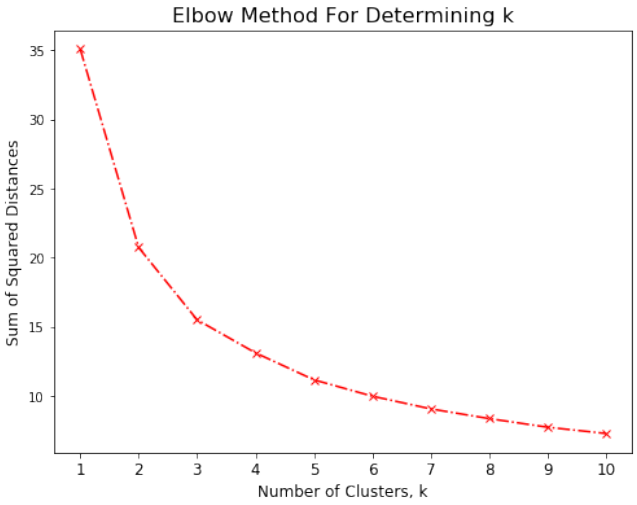
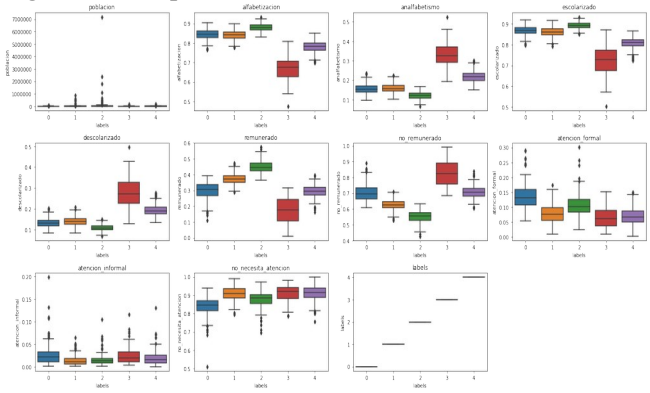


Figura 10: Boxplots of clusters before K-means



4 REFERENCE

1.https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018 [consultado 12/10/2019]