# Ecuador demographic analysis


- **Data scientist:** Yané, Ian Cristian Ariel
- **Project manager:** Mantilla, Jhon
- **Target company:** CORPORACIÓN GESTIÓN SOSTENIBLE

- **Dates (mm/dd/yyyy):**
  - **Initial date:** 08 / 23 / 2024
  - **Finish date:** 01 / 23 / 2024

## Summary:

### Brief Project Overview:

The project focuses on analyzing the distribution of the population by sector, age, and educational level with the aim of identifying sectors with a population over 50 years old in the province of Pichincha, specifically in the city of Quito. This information is crucial for future volunteer projects with an educational focus.

### Study Objectives:

- Population by Educational Level
- Population Distribution by Age and Educational Level
- Population by Area
- Population by Natural Region
- Population by Education Level and Sex
- Education Trends Between 2013 and 2014
- Population Age 60 Plus by approximately geographic Location

## Table of Contents:

# Introduction:

### Contextualization of the Problem:

In the context of a high proportion of older people experiencing loneliness and seeking companionship, especially those with education levels above the average (non-university higher education or tertiary education, university education, and postgraduate studies), and in contrast, the presence of numerous young individuals and people in general with a strong desire to learn but lacking the financial resources to afford education, the initiative of SUSTAINABLE MANAGEMENT CORPORATION arises. The proposal involves conducting an analysis to identify the necessary information for future volunteer projects, connecting older individuals willing to participate voluntarily with those seeking to acquire new skills.

# Stages:

In this data science study, Python 3.x will be used. After importing the necessary libraries, each stage follows the steps outlined below.

## Data Cleaning:

### Columns:

For this stage, I begin by decompressing and reading the .csv file with raw census information. After a review, unnecessary columns for the analysis, such as 'MONTO_VIVIENDA_MENSUAL', 'ZONA', and 'POSICION_OCUPACIONAL', were identified and removed.

Adjustments were made to column names for clarity and conciseness. For example, 'IDENTIF_SECT' was converted to 'ID_SECTOR', 'Ciudad Autorepresentada' was simplified to 'CIUDAD', 'Provincia' to 'PROVINCIA', and so on. This normalization facilitates data analysis and understanding.

Finally, the columns were rearranged hierarchically to improve the structure of the dataset. The new arrangement is as follows: 'ID_SECTOR', 'AREA_2000', 'PROVINCIA', 'ZONA_PLANIFICACION', 'REGION_NATURAL', 'SEXO', 'EDAD', 'ULTIMO_NIVEL_EDUCATIVO', 'GRADO_CURSO', 'ANIOS_TRABAJO', 'MES', 'ANIO', 'CIUDAD'.

### Null Values:

First, I replace null values, commonly represented as 'NaN', 'faltante', '-', among others within the datasets, with np.nan, the way the Pandas library interprets null values. Then, I use matrix and bar visualizations to assess the quantity and distribution of null values per column.

### Data Types:

I explore the data types present in the columns and observe that certain columns, such as 'AREA_2000', 'CIUDAD', 'PROVINCIA', 'REGION_NATURAL', 'SEXO', and 'ULTIMO_NIVEL_EDUCATIVO', contain categories. Therefore, I convert them to the 'category' data type for more efficient representation. Subsequently, I split the 'PERIODO' column into 'MES' and 'ANIO', converting them into integer columns along with the columns 'ZONA_PLANIFICACION', 'EDAD', 'GRADO_CURSO', and 'ANIOS_TRABAJO'.

### Standardization of Text Values to Lowercase:

I select the columns of type 'object' and 'category' to standardize text values, converting them to lowercase.

**Saving the clean DataFrame in .csv file:**

Finally, I save the clean .csv file for further exploratory data analysis.

## Exploratory Data Analysis (EDA):

**Setting Standard Parameters for Plots:**

I define the necessary parameters for the plots, ensuring visual consistency in all graphical representations.

**Special Functions Creation:**

**Encoding and Decoding Function:**

This function is responsible for encoding and decoding the dataframe as indicated by parameters. It also saves the column name and the instance used for encoding within a dictionary, allowing for later decoding if necessary. The function returns the encoded dataframe and the dictionary if encoding is specified, or the decoded dataframe if decoding is indicated.

**Mode Imputation Function:**

A function has been implemented to impute data by mode, as in the second iteration of the project it will be necessary to impute one of the target columns by this measure of dispersion. The function receives the dataframe and two columns, performs grouping by the first and the second, and returns the dataframe with null values within the second column imputed by the mode within the grouping by the first.

**Machine Learning Model Imputation Function:**

Given a sufficient amount of data within one of the target columns, a function has been created to impute data by a Machine Learning model. The function receives the encoded dataframe, the name of the column to impute, the convention for the null value within the encoded dataframe, and the local location of the model to use. Subsequently, it carries out the prediction process using the received parameters.

**Importing the clean .csv file into DataFrame:**

I start by importing the clean .csv file to continue the analysis.

**Exploratory Analysis:**

**Creation of Filtered DataFrame:**

I create a filtered dataframe with information about the target audience for the analysis. These are records of individuals who, as of now (12/2023), would be 50 years or older and belong to the province of Pichincha, city of Quito.

**Exploration with Seaborn:**

I use the Seaborn library to explore relationships between all columns through a pair plot. Additionally, I examine the frequency of variables in various columns ('EDAD', 'ULTIMO_NIVEL_EDUCATIVO', 'REGION_NATURAL', and 'AREA') using histograms.

**Correlation Matrix and Representative Pairs:**

I calculate a correlation matrix that shows the strength of the relationship between columns using the Pearson method. Subsequently, I identify those column pairs with correlations greater than +/- 70% and display them in a bar chart.

**Data Preprocessing:**

**Encoding and Scaling:**

I create a copy of the dataframe, assigning it to the variable 'codec_df'. Then, I encode its values using the 'label_encoder' function that I previously created. Later, I generate another copy of 'codec_df', assigning it to 'scaled_df', and scale the values using 'MinMaxScaler', adjusting them proportionally between 0 and 1.

**Imputation of Null Values:**

After completing the first iteration of the project, I perform the imputation of null values (np.nan). I start by imputing the values in the 'CIUDAD' column using the previously created 'impute_null_values_by_mode' function, which imputes based on the mode of cities within each province group. Subsequently, I impute the null values in the 'ULTIMO_NIVEL_EDUCATIVO' column using the model obtained during the Machine Learning stage, and display the number of missing values before and after imputation.

#### Saving codec and scaled DataFrame into .csv files:

I save the encoded, scaled, and imputed (in the second iteration) datasets for later use in the machine learning stage.

## Machine Learning:

**Creation of Special Functions:**

**grid_search_function:**

Finds the best parameters to train a machine learning model. Receives model parameters and test and train sets for X and y, returning a dictionary with the best parameters.

**train:**

Trains the model with the previously calculated best parameters. Receives the model and best parameters, and performs the training.

**test:**

Performs tests on the model and delivers accuracy.

**Importation of Encoded Dataset:**

I import the dataset previously encoded in the EDA stage and filter it to remove records with missing values in the columns of interest ('CIUDAD' and 'ULTIMO_NIVEL_EDUCATIVO').

**Data Separation into Features and Target:**

I separate the dataset into features (X) and the target column (y), with 'CIUDAD' in the first case and 'ULTIMO_NIVEL_EDUCATIVO' in the second. Then, I perform a split into training and test sets (X_train, y_train, X_test, and y_test).

**Regression Models:**

Two **regression models, linear and logistic**, were evaluated in both columns of interest ('CIUDAD' and 'ULTIMO_NIVEL_EDUCATIVO'). Both models did not provide optimal results and were discarded.

**Classification Models:**

**Support Vector Machine (SVM):**

Implemented the SVM model, yielding 1.0 (overfitting) in the 'CIUDAD' column and 0.31 in the 'ULTIMO_NIVEL_EDUCATIVO' column.

**Decision Tree:**

Used the decision tree model, achieving 1.0 (overfitting) in the 'CIUDAD' column and a quite good result of 0.76 in the 'ULTIMO_NIVEL_EDUCATIVO' column.

**Random Forest:**

The Random Forest model showed 1.0 (overfitting) in the 'CIUDAD' column and slightly outperformed the decision tree model with 0.78% in the 'ULTIMO_NIVEL_EDUCATIVO' column.

**K-Nearest Neighbors (KNN):**

Finally, implemented the KNN model, which gave 1.0 (overfitting) in the 'CIUDAD' column and a good performance of 0.67 in the 'ULTIMO_NIVEL_EDUCATIVO' column.

**Saving best model into .pfk file:**

For the 'CIUDAD' column, some models were discarded due to the limited amount of available data, resulting in either too low accuracy percentages or overfitting. For the 'ULTIMO_NIVEL_EDUCATIVO' column, the Random Forest model was stored as it provided a slightly higher percentage than the decision tree.

## Data Visualization:

**Relationship with Other Censuses:**

Below are two graphs illustrating the relationship between the 2010, 2014, and 2022 censuses in Ecuador. The first depicts the age distribution of the population in different censuses, while the second shows the percentage change in population by age between the censuses.

During the search for the dataset that would provide the most accurate information for the target objective, datasets from Ecuador's censuses in 2010, 2014, and 2022 were found. The decision was made to use the 2014 dataset because it required fewer inferences, providing values for the last educational level, ages, and country, province, and city of the individuals. Additionally, the comparison with the 2022 census showed

that the distribution of people by gender and age between the 2010 and 2022 censuses remained very similar, with barely perceptible variations.

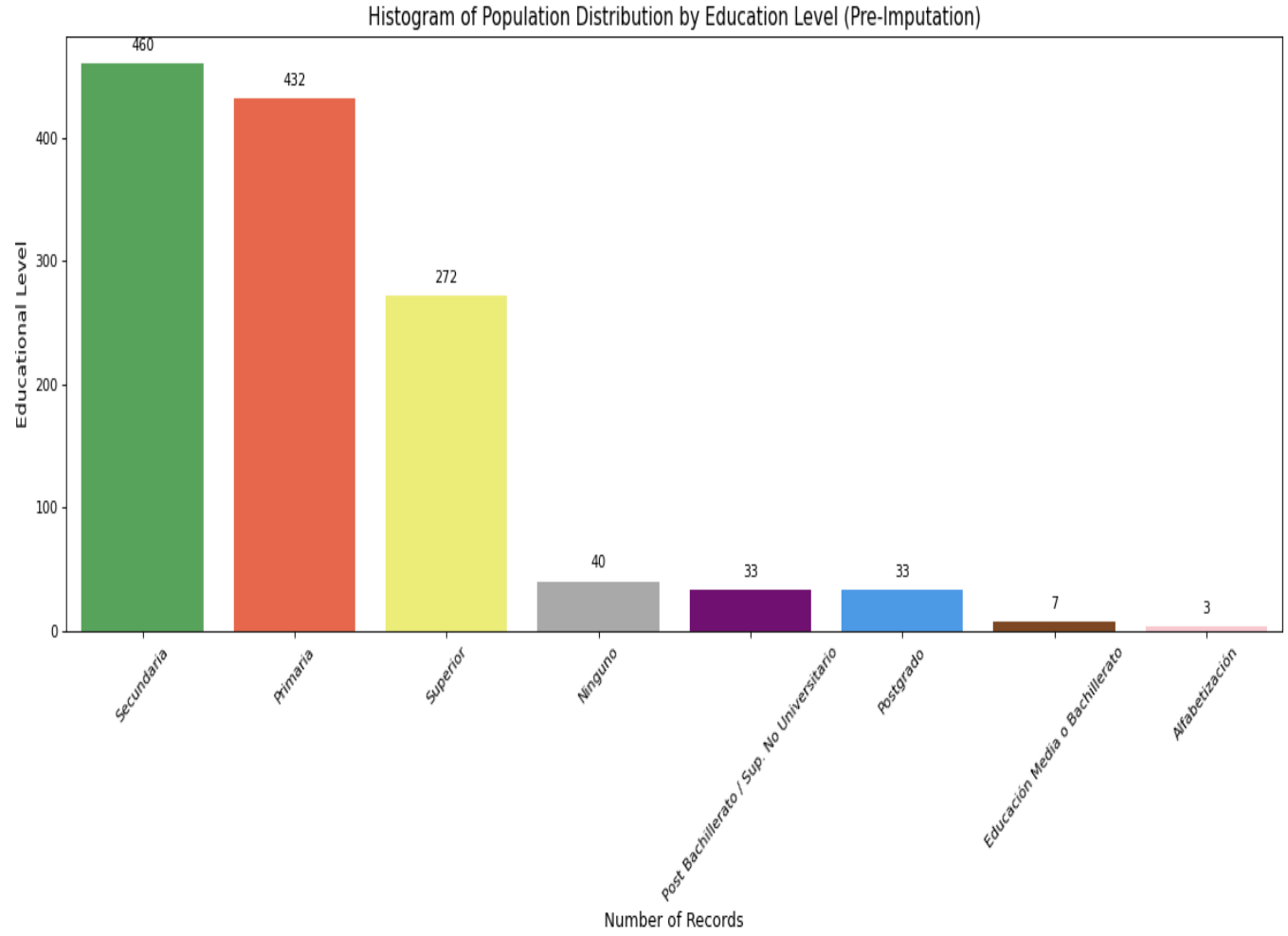
**Setting Parameters for Charts:**

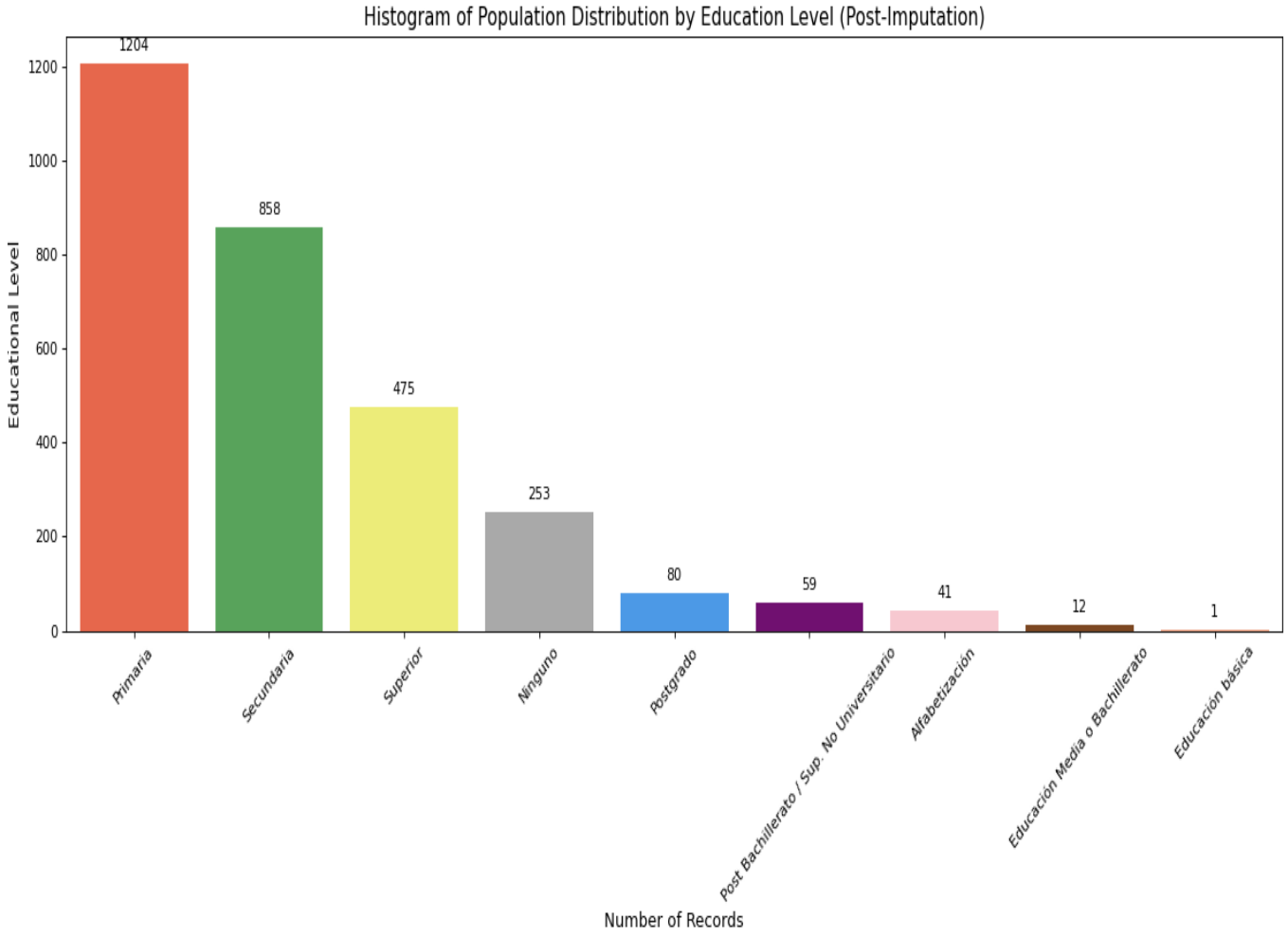
I define standard parameters for all my charts, including size, colors for target column variables, font color, among others.

**Loading before and after imputing .csv file into separate DataFrames:**

I load the dataset before imputation to observe the values prior to the imputation of values in the second iteration over age and the dataset that does have imputed values. Then, I filter each one for the target audience in terms of province (Pichincha), city (Quito), and age (individuals who, at present (2023), would be 50 years or older).

**Bar Chart of Population by Education Level:**


Histogram of Population Distribution by Education Level (Pre-Imputation)
Educational Level
Number of Records
460
432
272
40
33
33
7
3
Secundaria
Primaria
Superior
Ninguno
Post Bachillerato / Sup. No Universitario
Postgrado
Educación Media o Bachillerato
Alfabetización


Histogram of Population Distribution by Education Level (Post-Imputation)
Educational Level
Number of Records
1204
858
475
253
80
59
41
12
1
Primaria
Secundaria
Superior
Ninguno
Postgrado
Post Bachillerato / Sup. No Universitario
Alfabetización
Educación Media o Bachillerato
Educación básica

**Before Imputation Conclusion:** The main education levels in the analysis are:

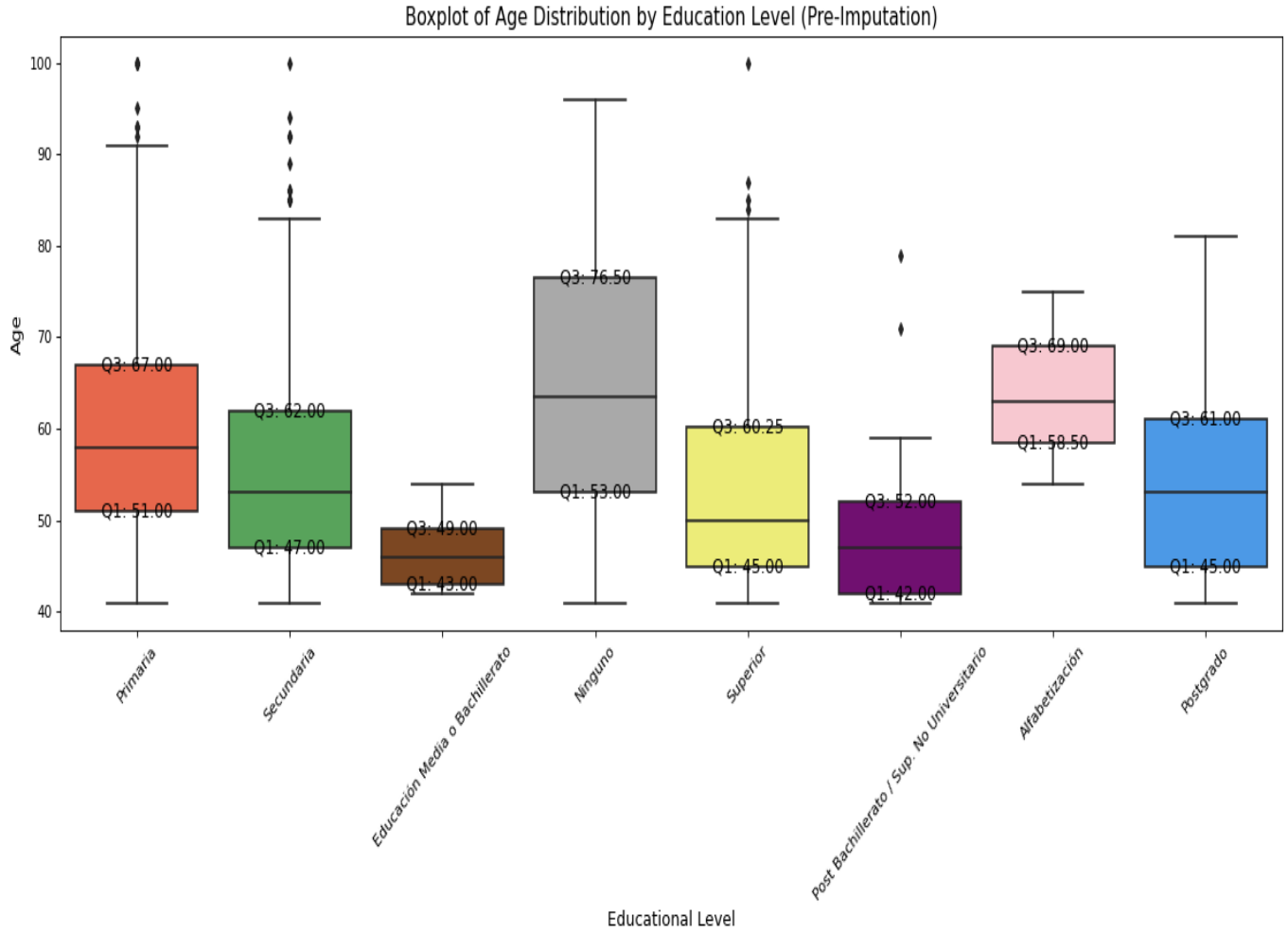
- Secondary with 858 records.
- Primary with 432 records.
- Higher with 272 records.
- Columns with fewer records (none, post high school, postgraduate, bachelor's degree, and literacy skills) have less presence.

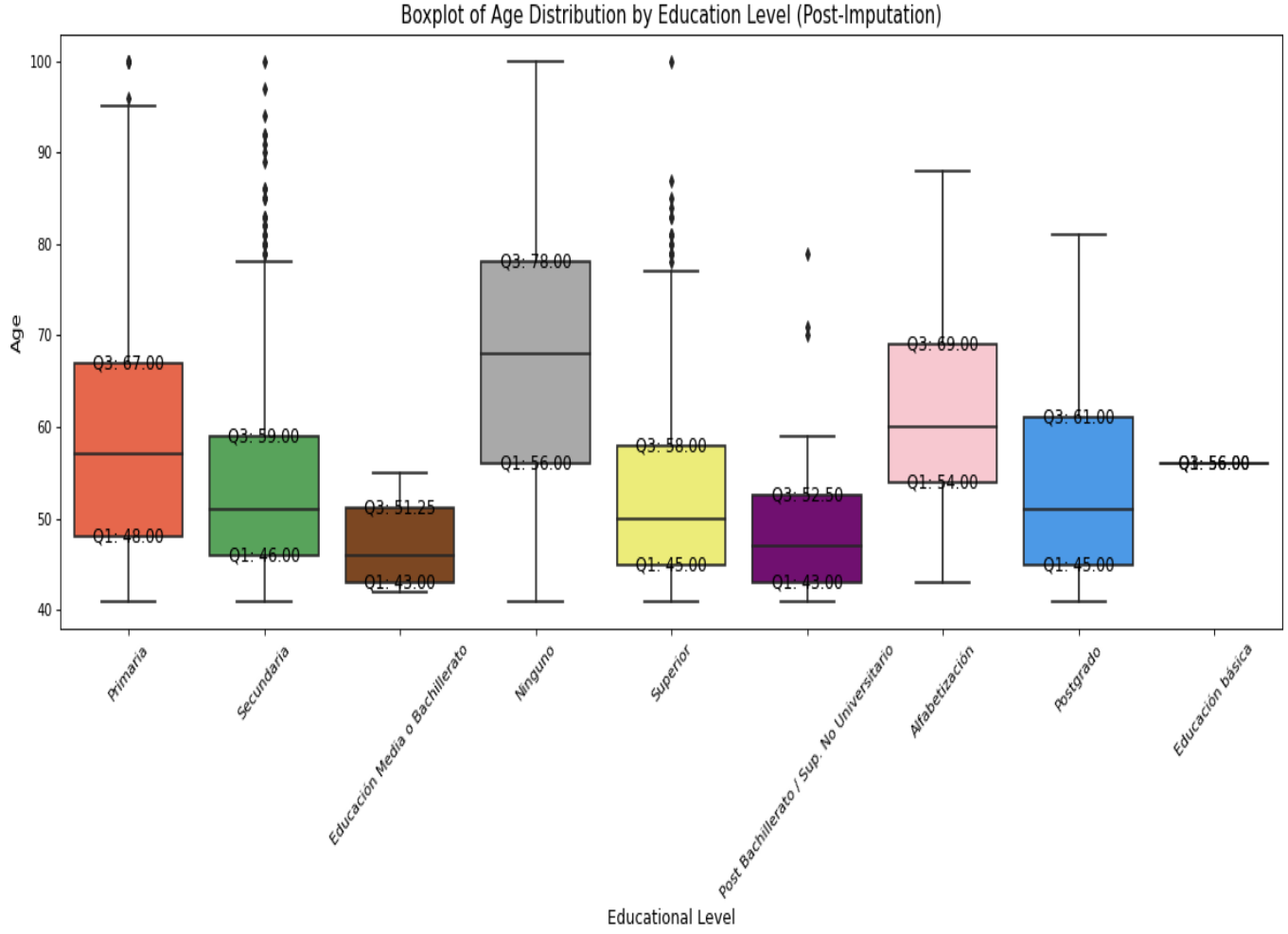
**After Imputation Conclusion:** The main education levels in the analysis are:

- Primary with 1204 records.
- Secondary with 858 records.
- Higher with 475 records.
- Columns with fewer records (none, postgraduate, post high school, literacy, and bachelor's degree) have less presence.

This analysis provides a clear view of the population distribution based on their educational level before and after data imputation. There is a shift in the frequency of records in each education level after imputation, which may influence future decisions or actions related to demographic analysis.

**Boxplot of Ages by Education Levels:**


Boxplot of Age Distribution by Education Level (Pre-Imputation)
Age
Educational Level
Primaria
Secundaria
Educación Media o Bachillerato
Ninguno
Superior
Post Bachillerato / Sup. No Universitario
Alfabetización
Postgrado
Q3: 67.00
Q1: 51.00
Q3: 62.00
Q1: 47.00
Q3: 49.00
Q1: 43.00
Q3: 76.50
Q1: 53.00
Q3: 60.25
Q1: 45.00
Q3: 52.00
Q1: 42.00
Q3: 69.00
Q1: 58.50
Q3: 61.00
Q1: 45.00


Boxplot of Age Distribution by Education Level (Post-Imputation)
Age
Educational Level
Primaria
Secundaria
Educación Media o Bachillerato
Ninguno
Superior
Post Bachillerato / Sup. No Universitario
Alfabetización
Postgrado
Educación básica
Q3: 67.00
Q1: 48.00
Q3: 59.00
Q1: 46.00
Q3: 51.25
Q1: 43.00
Q3: 78.00
Q1: 56.00
Q3: 58.00
Q1: 45.00
Q3: 52.50
Q1: 43.00
Q3: 69.00
Q1: 54.00
Q3: 61.00
Q1: 45.00
Q3: 56.00
Q1: 56.00

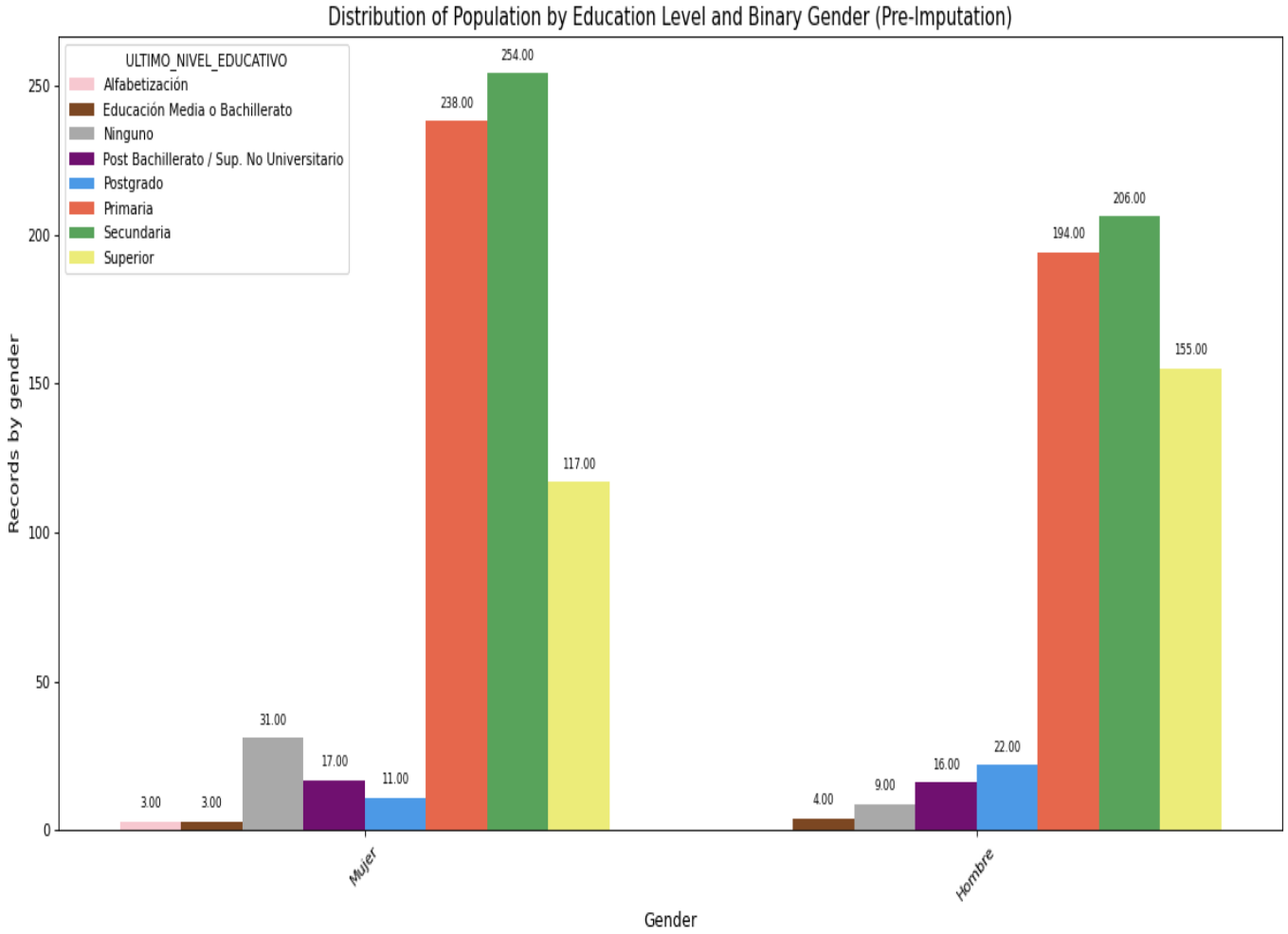
**Before Imputation Conclusion:**

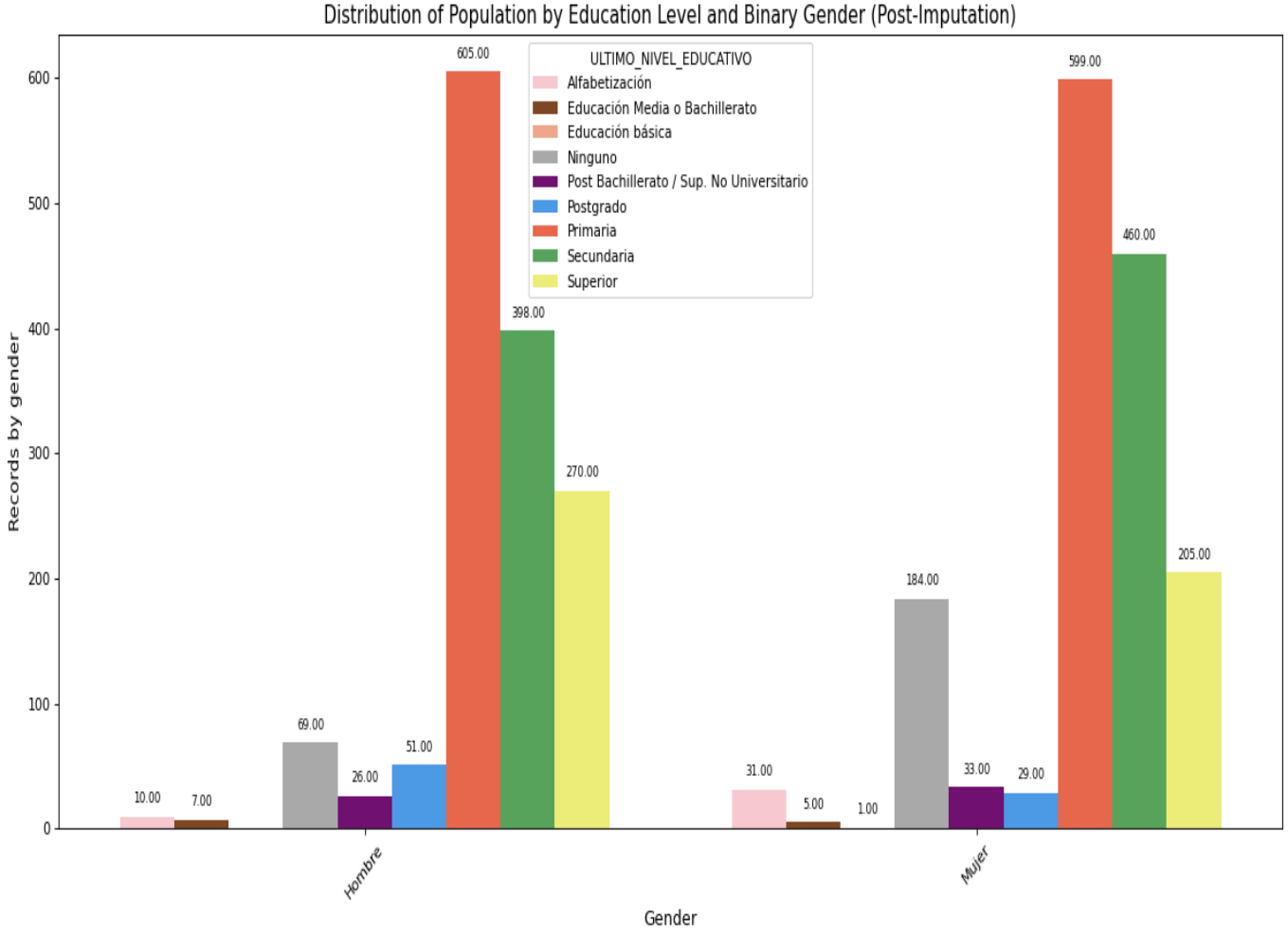
- The age range for individuals with no formal education spans between 53 and 76 years.
- For those with literacy skills, the age range is observed between 50 and 69 years.
- For individuals with primary school education, the highest number of records is observed between 51 and 67 years old.
- Those with secondary education show a concentration between 47 and 62 years.
- The bachelor's degree category shows its peak between 43 and 49 years.
- Individuals with a higher education background are most prevalent between 45 and 60 years.
- Individuals with post-bachelor's or non-university higher education typically fall between 42 and 52 years.
- Individuals pursuing postgraduate education tend to be between 45 and 61 years old.

**After Imputation Conclusion:**

- The age range for individuals with no formal education spans between 56 and 70 years.
- For those with literacy skills, the age range is observed between 54 and 69 years.
- For individuals with primary school education, the highest number of records is observed between 40 and 67 years old.
- Those with secondary education show a concentration between 46 and 59 years.
- The bachelor's degree category shows its peak between 43 and 51 years.
- Individuals with a higher education background are most prevalent between 45 and 50 years.
- Individuals with post-bachelor's or non-university higher education typically fall between 43 and 52 years.
- Individuals pursuing postgraduate education tend to be between 45 and 61 years old.

**Bar Chart by Gender for Education Level:**


Distribution of Population by Education Level and Binary Gender (Pre-Imputation)
ULTIMO_NIVEL_EDUCATIVO
Alfabetización
Educación Media o Bachillerato
Ninguno
Post Bachillerato / Sup. No Universitario
Postgrado
Primaria
Secundaria
Superior
Records by gender
Gender
Mujer
Hombre
3.00
3.00
31.00
17.00
11.00
238.00
254.00
117.00
4.00
9.00
16.00
22.00
194.00
206.00
155.00


Distribution of Population by Education Level and Binary Gender (Post-Imputation)
ULTIMO_NIVEL_EDUCATIVO
Alfabetización
Educación Media o Bachillerato
Educación básica
Ninguno
Post Bachillerato / Sup. No Universitario
Postgrado
Primaria
Secundaria
Superior
Records by gender
Gender
Hombre
Mujer
10.00
7.00
69.00
26.00
51.00
605.00
398.00
270.00
31.00
5.00
1.00
184.00
33.00
29.00
599.00
460.00
205.00

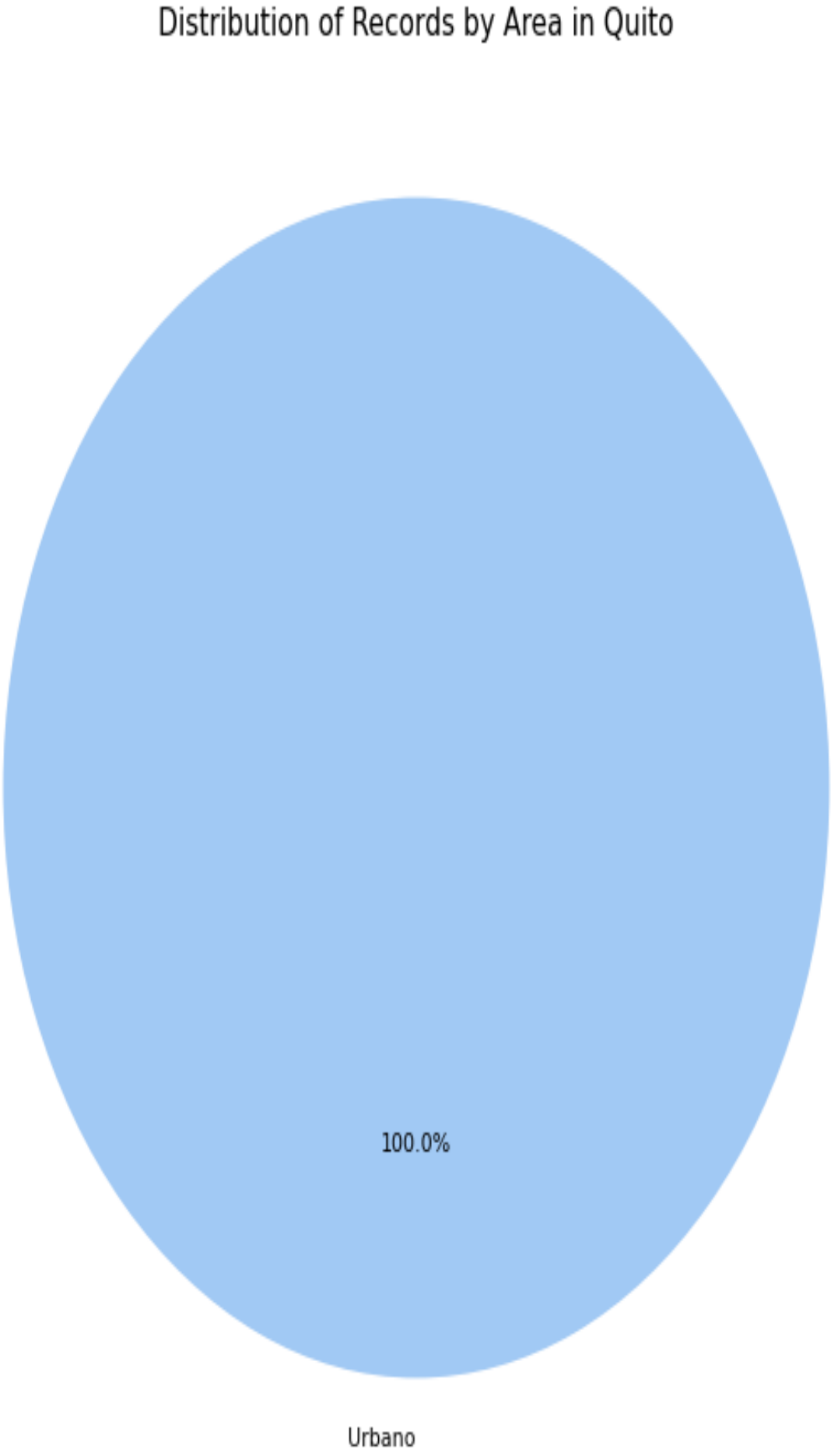
**Before Imputation Conclusion:**

- In general, it is observed that females tend to have their highest academic attainment in categories such as none, primary, higher, and secondary education compared to males.
- Conversely, males exhibit a higher prevalence in the advanced education degrees, for example, postgraduate and higher education categories.

**After Imputation Conclusion:**

- The same differences remain as before, indicating consistency in gender-based educational disparities.

**Pie Charts for Area:**


Distribution of Records by Area in Quito
100.0%
Urbano

**Before Imputation:**

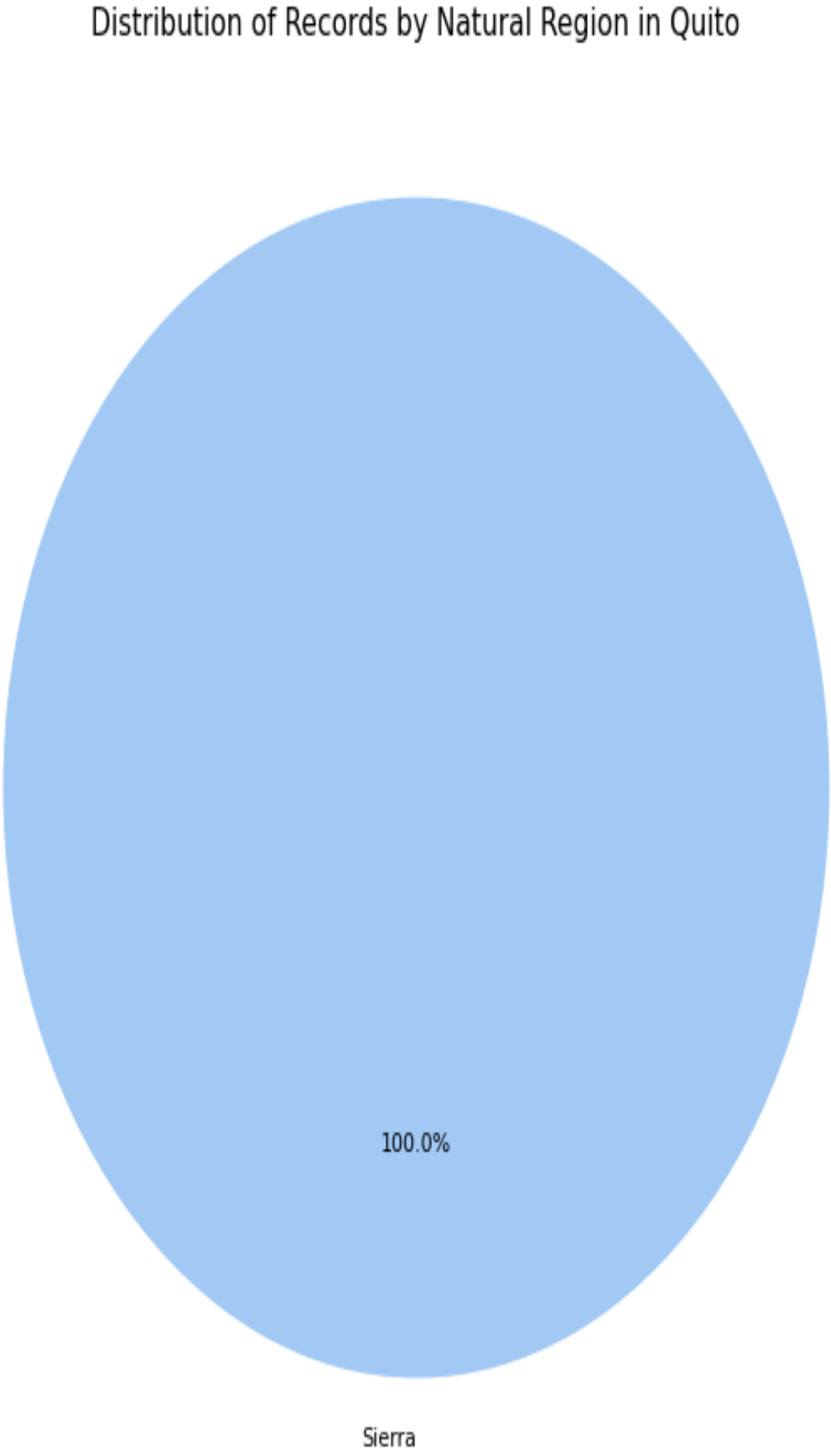
- The data reveals that the area is 100% urban.

Note: As the entire area is urban, no further analysis related to the area will be derived.

**After Imputation:**

- No added new information after imputing null values in the column. This analysis indicates that both before and after imputation, the entirety of the area is classified as urban.

**Pie Charts for natural region:**


Distribution of Records by Natural Region in Quito
100.0%
Sierra

**Before imputation:**

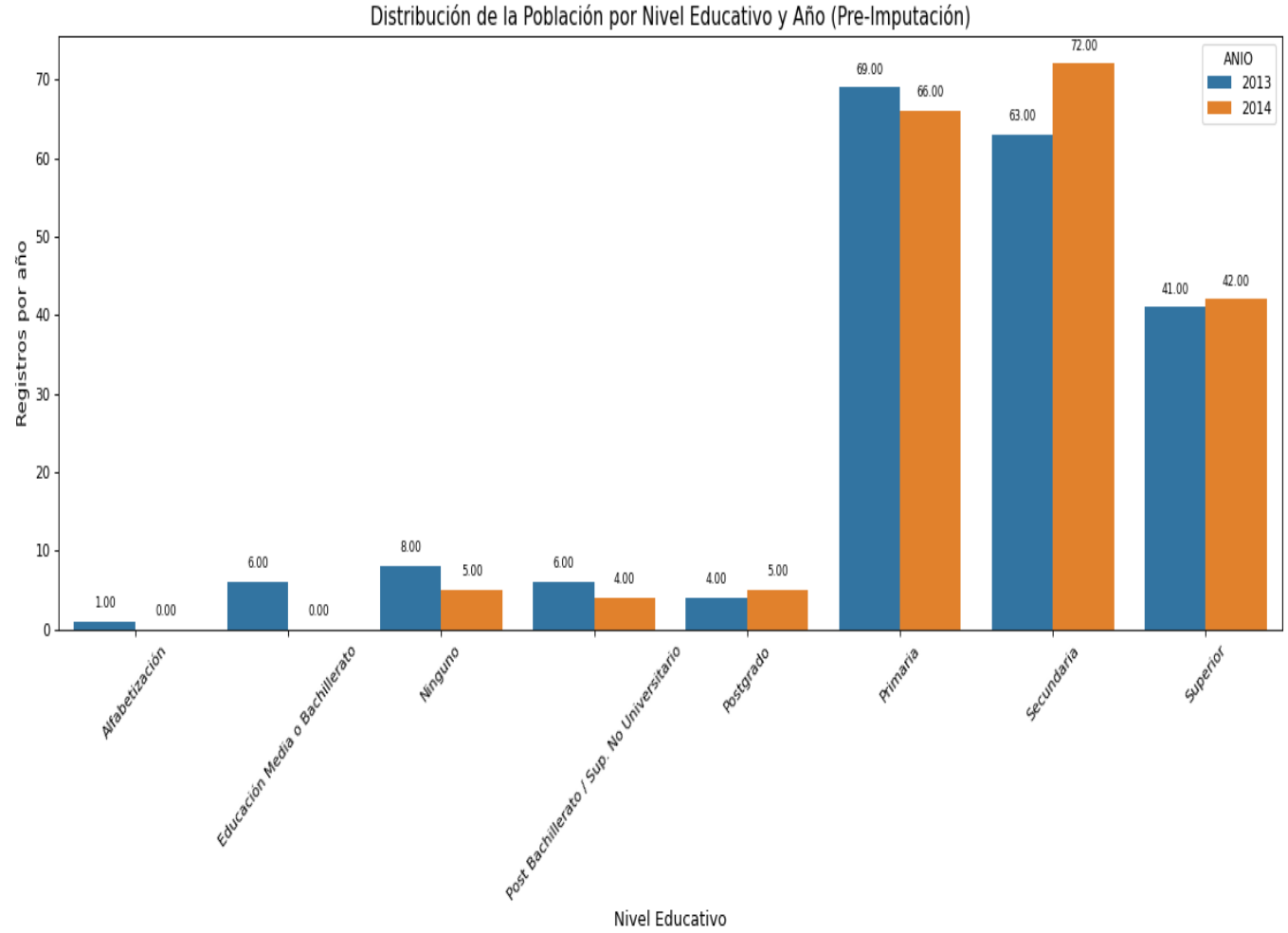
- The data indicates that the natural region is **100% Sierra**.

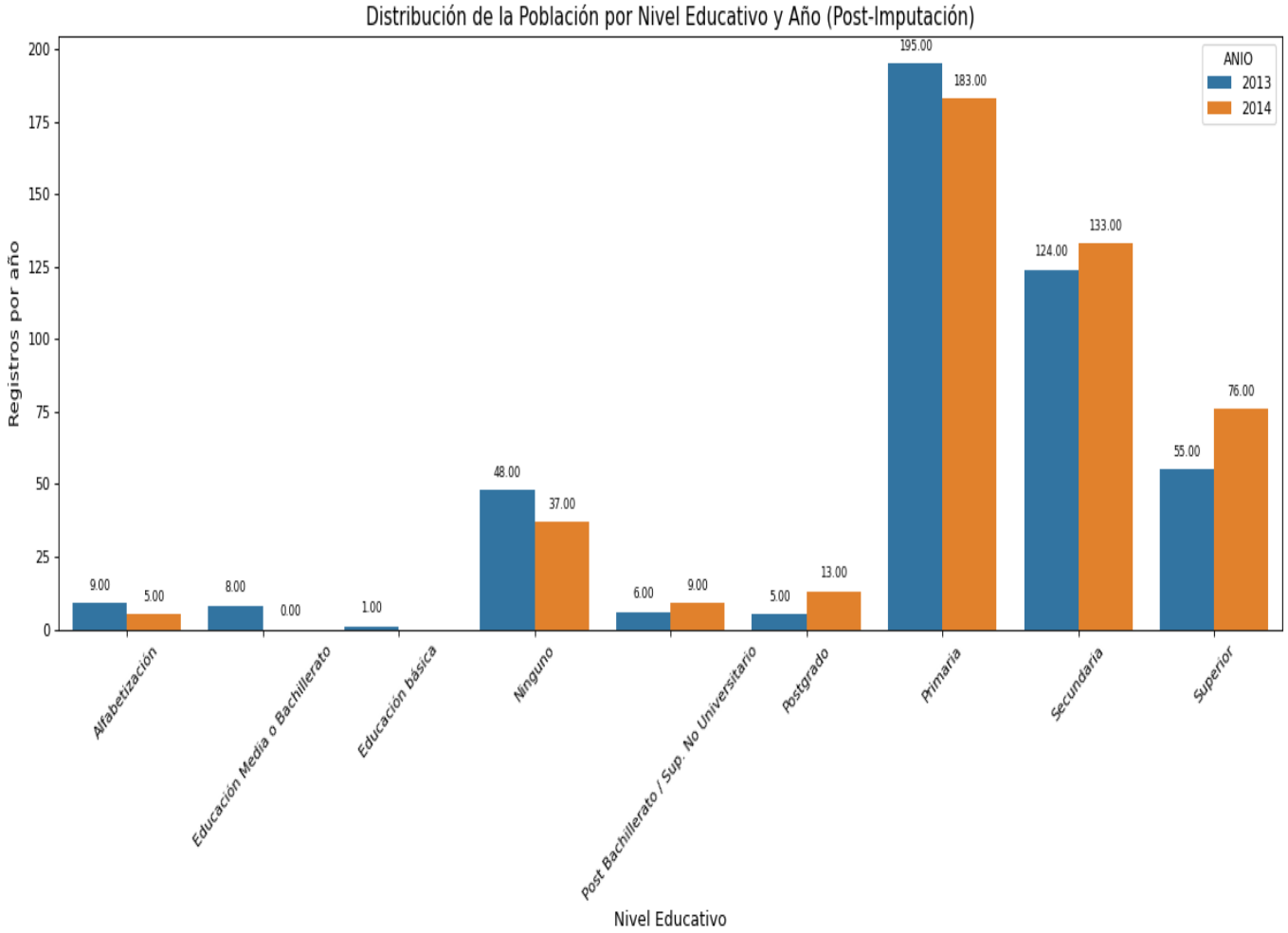
**Note:** As the entire region is Sierra, no further analysis related to the natural region will be derived.

**After imputation:**

- no added new information later imputation null values in column

**Bar Charts by Education Level According to Years:**


Distribución de la Población por Nivel Educativo y Año (Pre-Imputación)
Registros por año
Nivel Educativo
ANIO
2013
2014
Alfabetización
Educación Media o Bachillerato
Ninguno
Post Bachillerato / Sup. No Universitario
Postgrado
Primaria
Secundaria
Superior
1.00
0.00
6.00
0.00
8.00
5.00
6.00
4.00
4.00
5.00
69.00
66.00
63.00
72.00
41.00
42.00


Distribución de la Población por Nivel Educativo y Año (Post-Imputación)
Registros por año
Nivel Educativo
ANIO
2013
2014
Alfabetización
Educación Media o Bachillerato
Educación básica
Ninguno
Post Bachillerato / Sup. No Universitario
Postgrado
Primaria
Secundaria
Superior
9.00
5.00
8.00
0.00
1.00
48.00
37.00
6.00
9.00
5.00
13.00
195.00
183.00
124.00
133.00
55.00
76.00

**Before Imputation:**

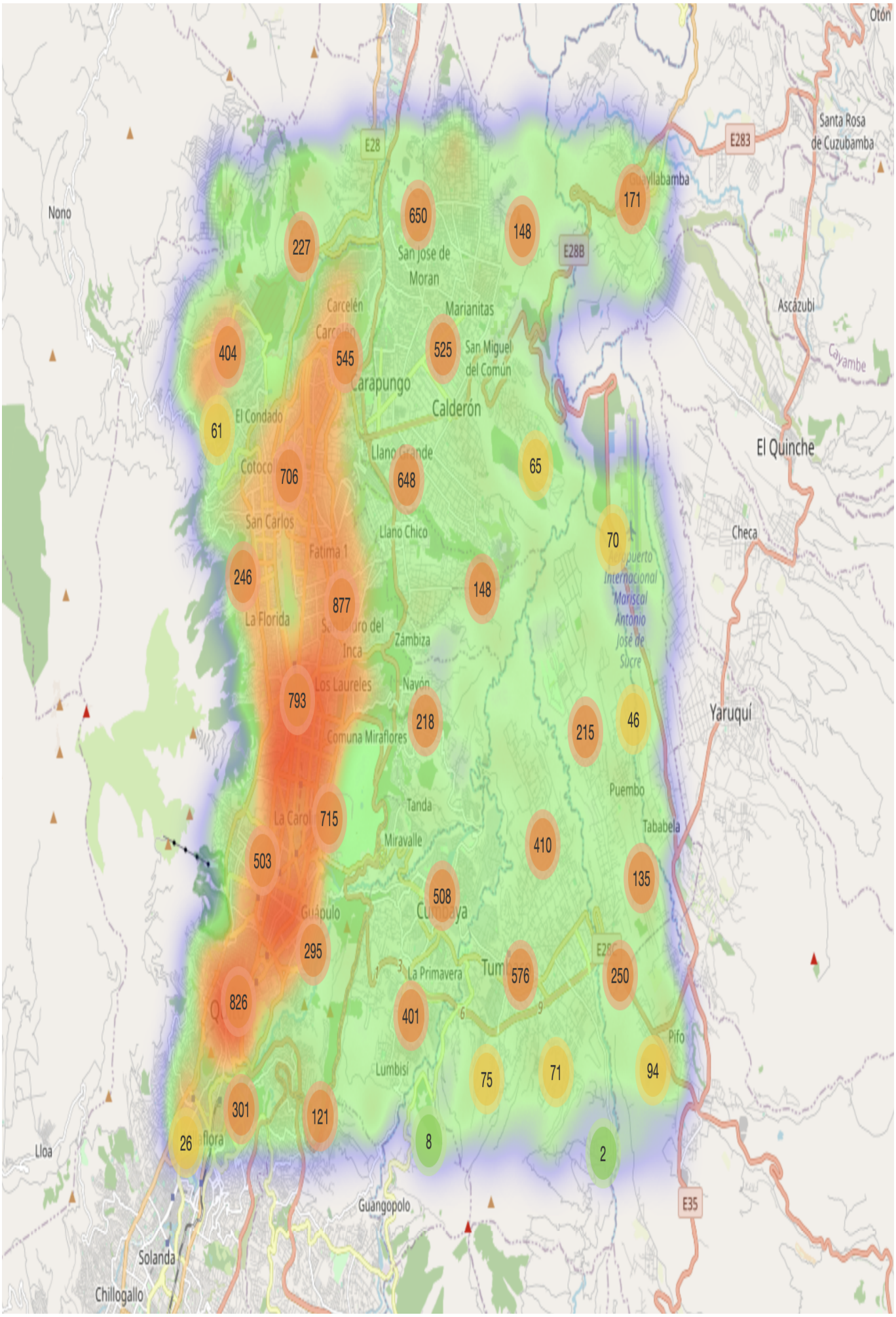
- The analysis reveals that in categories such as literacy, no formal education, non-university higher education, and primary, there is a mild increase in the number of records in the year 2013.
- Interestingly, in 2014, there is a subtle increase in the index for remaining categories like secondary, higher education, and postgraduate education.

Note: This indicates variations in educational distribution between slightly different categories between the years 2013 and 2014.

**After Imputation:**

- The overall observations remain consistent before imputation, except for the Non-university Higher Education category, which shows a slight increase in the year 2014 in this case.
- This analysis highlights mild changes in educational distribution between the years 2013 and 2014, which could be of interest when planning educational strategies or assessing trends over time.

## Interactive Map with Heatmap for People Aged 65 and Older in Ecuador, Pichincha, Quito (2020):


Otón
Santa Rosa de Cuzubamba
E283
E28
E28B
Nono
227
650
San jose de Moran
148
171
Marianitas
Carcelén
404
545
525
San Miguel del Común
Carapungo
Calderón
Ascázubi
Cayambe
61
El Condado
Cotoco
706
Llano Grande
648
65
El Quinche
San Carlos
Llano Chico
70
Checa
Fatima 1
246
877
148
La Florida
Zámbiza
Inca
Los Laureles
Nayón
793
218
215
46
Yaruquí
Comuna Miraflores
Puembo
Tanda
715
Miravalle
Tababela
410
503
135
508
Guápulo
295
La Primavera
576
250
826
401
Pifo
Lumbisí
75
71
94
301
121
26
8
2
Lloa
Guangopolo
E35
Solanda
Chillogallo

The population aged 60 and above is distributed along the main streets in the following areas:

**Lengthwise Distribution:**

- Coordinates at the start:
  - Latitude: -0.232127
  - Longitude: -78.5130033
- Coordinates at the end:
  - Latitude: -0.077186
  - Longitude: -78.464020

**Widthwise Distribution:**

- Coordinatesat the start:
  - Latitude: -0.148391
  - Longitude: -78.502753
- Coordinates at the end:
  - Latitude: -0.1595376
  - Longitude: -78.4577653

This map provides an effective visualization of the concentration of the population aged 65 and older in specific areas of Ecuador, Pichincha, Quito during the year 2020. Geospatial information can be valuable for planning services and resources targeted at the elderly population in those locations.

# Discussion:

## Relating Findings to Objectives:

The specific results of the analysis, such as the location, gender, and types of educational levels of the target audience, were observed.

# Conclusions:

## Summary of the Most Important Results:

In this analysis, data from the 2013-2014 census was used as it was the richest dataset to obtain information about the educational level of the population. Additionally, after investigating population variability until the latest census in 2022, a similar population distribution by age is observed, as detailed in the census report provided by www.censoecuador.gob.ec (pages 29 to 31).

The population aged 60 and above is distributed along the principal streets in the following areas:

**Lengthwise Distribution:**

- Initial coordinates:
  - Latitude: -0.232127
  - Longitude: -78.5130033
- Finish coordinates:
  - Latitude: -0.077186
  - Longitude: -78.464020.

**Widthwise Distribution:**

- Initial coordinates:
  - Latitude: -0.148391
  - Longitude: -78.502753
- Finish coordinates:
  - Latitude: -0.1595376
  - Longitude: -78.4577653.

**Non-University Higher Education:**

- Age range: 43 to 52 years.
- Mainly pronounced on females.

**Higher Education:**

- Age range: 45 to 50 years.
- Mainly pronounced on males.

**Postgraduate Education:**

- Age range: 45 to 61 years.
- Mainly pronounced on males.

# Recommendations:

## To future research:

The dataset used in this analysis adequately addressed all posed questions. For future research, it is suggested to consider the use of updated census datasets. If multiple datasets are needed to gather the necessary information, following appropriate inference steps, similar to those performed in this analysis by adjusting census data to the current date and supporting it with statistical and official sources, is crucial.

## To apply this research:

It is recommended to review this report in detail to understand the context, result accuracy, and the potential for scaling or reprofiling the analysis to other locations in Ecuador. Additionally, updating the analysis with new data based on evolving demographic trends is advisable.

# Bibliography:

## List of All Cited Sources in the Report:


- ChatGPT 3.5
- Encuesta de Condiciones de Vida - ECV
- Población de 65 años de edad y más (% del total)
- 2022, Ecuador: Informacion censal
- Población, total - Ecuador

- Ecuador - Subnational Demographic and Health Data

# Appendices:

### Additional Details:

- Clean data stage
- Exploratory data analisis stage
- Machine learning stage
- Data visualization stage