Using Python to Analyse the Bulgarian Census Years: What the Numbers Tell Us about the Roma Population

by

Ivan Ivanov

Submitted to

Central European University

Roma Graduate Preparation Program

Supervisors: Pal Galicza & Mihaly Orsos

Budapest, Hungary

2024

Copyright notice

I, the undersigned Ivan Ivanov, hereby declare that I am the sole author of this thesis. This thesis contains no material previously published by any other person except where proper acknowledgment has been made. This thesis contains no material which has been accepted as part of the requirements of any other academic degree or nondegree program, in English or in any other language. This is a true copy of the thesis, including the final version.

Date: 7th June 2024

Name (printed): Ivan Ivanov

Signature: ИВАН ИВАНОВ

Abstract

This paper delves into the analysis of the Bulgarian census data with a focus on understanding the dynamics of the Roma population. Censuses play a pivotal role in government planning and policy formulation by providing essential demographic, educational, and socioeconomic information about the population. The history of census-taking in Bulgaria reflects evolving societal norms and international standards, with recent censuses incorporating voluntary inquiries on ethnicity, language, and religion to ensure a more accurate representation of the country's diverse population. It presents an analysis of the Roma population trends in Bulgaria using linear regression. By examining census data from 1900 to 2021, the paper predicted the Roma population for the year 1975, which was excluded from the training data. The results show that the linear regression model closely approximates the actual population, demonstrating the effectiveness of this approach for time-series prediction. The findings indicate disparities between the overall Bulgarian and the Roma populations regarding age distribution, educational attainment and labour force participation, emphasizing the importance of comprehensive strategies to address barriers to employment and foster inclusive economic participation. This analysis underscores the imperative of understanding and addressing the diverse needs of Bulgaria's multiethnic population. By leveraging insights from census data and employing evidence-based approaches, policymakers can work towards building a more just, equitable, and prosperous future for all Bulgarians.

Key words: Bulgaria, Roma, Census, Population, Analyse

Table of Contents

CHAPTER 1: Introduction

Censuses are important for governments all around the world because they are not just about counting people but provide relevant economic and social information. According to the definition by Baffour et al. (2013), it "provides important information on a country's population that is used in government planning and to underpin the national statistical system" (p. 337). First, a census gives an accurate number of people living in an area. This helps governments plan and decide how to distribute resources like money and services, such as schools and hospitals. Furthermore, it helps determine how many representatives each place gets in the government. Second, the census collects information about different aspects of people's lives, such their age, gender, race, and family. This enables governments to understand what their society looks like and how it changes over time.

Another important aspect is that the census decides how to share federal money among different areas. This money is crucial for financial decisions, such as building schools, hospitals, and roads. So, it is important to make sure everyone gets a fair share. Moreover, the census helps governments, businesses, and organizations make plans for the future. By knowing where people live and what they need, they can decide where to put new schools, hospitals, or businesses. Lastly, the census data is used by researchers and policymakers to study several issues, for example how populations change, how economies grow, and how to keep people healthy and safe. The census is not just about counting people. It is an important tool for understanding society, in order to make fair decisions, and share resources equally for everyone's benefit (Baffour et al., 2013).

The focus on ethno-cultural aspects of the Bulgarian population has been a key aspect since the inception of census-taking, which spans a history of 135 years in Bulgaria. Initially, during the late 19th century, characteristics such as 'religion' and 'native language' were documented in the censuses of 1887 and 1892. Subsequently, in 1900, 'nationality' was added as a characteristic. In subsequent censuses until 1975, 'mother tongue' replaced 'native language,' while 'nationality' remained a significant factor. However, 'religion' was omitted from consideration after the 1946 census. The characteristics 'ethnic group,' 'mother tongue,' and 'religion' were reintroduced in the censuses of 1992 and 2001. Then, in the censuses of 2011 and 2021, a new characteristic, 'religion,' was included. It is essential to note that in the last three censuses, inquiries concerning these ethno-cultural aspects were voluntary, aligning with international standards established by the United Nations. These standards underscore principles such as voluntary participation, self-determination, and the freedom to record answers based on personal identification (Bulgarian census, 2021).

Population studies are crucial for understanding demographic changes and planning social policies. The Roma population in Bulgaria has experienced significant changes over the past century. This paper aims to predict the Roma population for the year 1975 using historical census data and linear regression. The prediction's accuracy is evaluated by comparing the model's output to the actual census data for 1975.

A crucial tool in analysing census data is Python, a powerful programming language, which offers a diverse set of tools and libraries that assist in various aspects of data processing and analysis. Firstly, Python facilitates data collection from different sources, including census databases and online repositories. Tools like Pandas help clean and prepare the data for analysis. Furthermore, Python enables data exploration and visualization through libraries such as Matplotlib, Seaborn, and Plotly. These tools create insightful visualizations like charts, graphs, and aps, making it easier to interpret complex demographic trends within census data. Additionally, Python provides libraries like NumPy and SciPy for statistical analysis, allowing researchers to uncover relationships, trends, and correlations within the data. Machine learning algorithms from the scikit-learn library can be applied to census data for predictive modelling and identifying patterns within the population. For census data containing geographic information, Python libraries like GeoPandas, Shapely, and Folium facilitate geospatial analysis and mapping, enabling researchers to visualize spatial patterns and distributions across different regions. Python's web scraping capabilities, with libraries like Requests and Beautiful Soup, enable the extraction of census data from websites. Moreover, Python can interact with APIs provided by census bureaus for real-time data updates and integration into analysis workflows. Overall, Python offers a versatile and comprehensive ecosystem of tools and libraries that greatly aid in understanding and analyzing census data. Its ease of use, extensive documentation, and active community support make it an ideal choice for data professionals and researchers working with census data.

CHAPTER 2: Aspects of Analysing Census Data

The Roma population, a historically marginalized ethnic group in Europe, has faced significant challenges in demographic documentation and socio-economic integration. The diversity within the Roma community and various identification practices complicates accurate population estimation. This review synthesizes data from Bulgarian censuses and other sources to understand the Roma demographic trends from 1900 to 2021, emphasizing the impact of socio-political changes on population reporting, educational attainment, labor force participation, and linguistic diversity.

2.1. Historical Population Estimates

The Roma population in Bulgaria has been inconsistently recorded due to political influences and identification practices. Under the communist regime, the number of Roma was significantly underreported. For instance, the 1975 census recorded only 18,000 Roma, while a Ministry of the Interior survey in 1980 estimated their number at 524,000, highlighting discrepancies due to different methodologies (Bulgarian census, 2021). By 1989, the Ministry estimated approximately 577,000 Roma.

Post-1990, estimates have varied widely. The Library of Congress reported around 450,000 Roma in 1990, while the United Nations Development Program (UNDP) estimated between 600,000 and 750,000 at the beginning of the 21st century. Recent census data, however, show a declining trend in self-identified Roma, with 320,761 in 2011 and 266,720 in 2021, reflecting changes in self-identification and demographic factors (Bulgarian census, 2021).

2.2. Data Analysis and Model Training

A study utilized census data from 1900 to 2021, excluding 1975, to train a Linear Regression model predicting the Roma population. The model is using the R^2 (coefficient of determination) value provides a intuitive measure of how well the model explains the variance in the data. The linear regression model accurately predicted the Roma population for the year 1975, closely matching the real population, not the counted one.

2.3. Linguistic Diversity

Linguistic data from the 2021 census revealed significant numbers of Bulgarian and Turkish speakers, with Romani speakers constituting a smaller segment. This linguistic distribution underscores the diverse ethnic composition of Bulgaria and highlights the cultural presence of the Roma community despite their smaller numbers (Bulgarian census, 2021).

2.4. Long-Term Impacts and Legacy

The 1975 census and its ensuing policy measures represent a watershed moment, encapsulating the intricate dynamics between the state and the Roma population in Bulgaria. The transition from nomadic to settled lifestyles aimed at improving living standards brought significant disruptions to traditional Roma ways of life. While urbanization offered access to modern amenities and services, it also led to the erosion of cultural practices deeply rooted in Roma heritage.

The educational and vocational initiatives presented both opportunities and challenges. Access to formal education and vocational training promised socio-economic advancement but often overlooked the cultural and linguistic needs of Roma students, perpetuating disparities in educational outcomes and employment opportunities. The promotion of Bulgarian language and culture marginalized Roma cultural identities, exacerbating feelings of alienation and exclusion.

Economically, the assimilationist policies had profound repercussions. Employment opportunities within state enterprises and cooperatives offered a semblance of stability for some Roma individuals. However, broader economic restructuring under socialism disproportionately impacted marginalized communities, perpetuating cycles of poverty and marginalization.

2.5. Age Distribution

Age distribution data indicate a younger demographic profile for the Roma compared to Bulgarians. The largest cohorts among Bulgarians were in the working age groups (30-59), while the Roma had substantial numbers in younger age brackets (0-19), suggesting implications for social services and education tailored to a younger population (Bulgarian census, 2021).

2.6. Educational Attainment

Educational disparities between the Roma and Bulgarian populations are pronounced. Roma individuals exhibit lower levels of educational attainment, with a small proportion achieving higher education compared to Bulgarians. This gap underscores the need for targeted educational interventions to address socio-economic barriers and promote inclusivity (Bulgarian census, 2021).

2.7. Labor Force Participation

The labor force data from 2021 reveal significant disparities. While a considerable number of Bulgarians are employed, the Roma community faces higher unemployment and lower participation rates. Addressing these disparities requires targeted policies to enhance employment opportunities and socio-economic integration for the Roma (Bulgarian census, 2021).

CHAPTER 3: Methodology

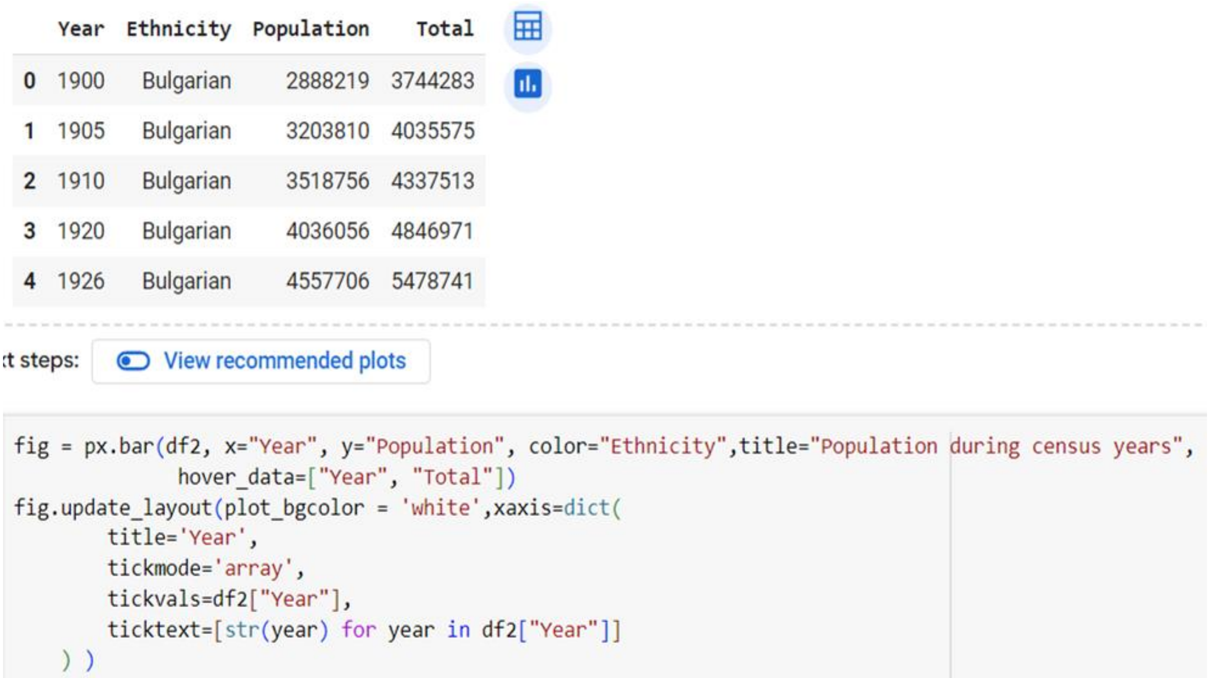
This methodology employs Python for creating visual representations of data and performing linear regression analysis. Python offers a robust ecosystem for data analysis, with libraries such as Plotly Express for chart creation and scikit-learn for machine learning, including linear regression. This approach allows for comprehensive data visualization and predictive modelling, making it easier to understand trends and make informed predictions.

3.1. Data Visualization with Plotly Express

Plotly Express is a high-level data visualization library in Python, which makes it simple to create interactive and visually appealing charts. Below, it is illustrated how to generate a bar chart using Plotly Express. This example will visualize population data over different years, with bars coloured by ethnicity, and includes hover functionality to display additional information about each bar. The layout is customized to enhance readability and aesthetics (Figure 1).

Figure 1

Python session in Google Colab



	Year	Ethnicity	Population	Total
0	1900	Bulgarian	2888219	3744283
1	1905	Bulgarian	3203810	4035575
2	1910	Bulgarian	3518756	4337513
3	1920	Bulgarian	4036056	4846971
4	1926	Bulgarian	4557706	5478741

```
fig = px.bar(df2, x="Year", y="Population", color="Ethnicity",title="Population during census years",
             hover_data=["Year", "Total"])
fig.update_layout(plot_bgcolor = 'white',xaxis=dict(
        title='Year',
        tickmode='array',
        tickvals=df2["Year"],
        ticktext=[str(year) for year in df2["Year"]]
    ) )
```

3.2. Linear Regression Analysis with Scikit-Learn

Linear regression is a fundamental technique in predictive modelling. Using the scikit-learn library, we can fit a linear model to our data to predict unknown values. Here, I illustrate the process of training a machine learning model to predict population data and evaluate its performance.

3.3. Data Retrieval and Preparation

First, the dataset is loaded, which includes the population data and the total population of a given region over a series of years.

3.4. Data Splitting

Divide the data into training and testing sets. The training set is used to train the model, while the testing set evaluates its performance.

3.5. Model Training

A linear regression model is trained on the training data. The model learns the relationship between the independent variable (year) and the dependent variable (population).

3.6. Model Evaluation

The model's performance is assessed using metrics such as R^2 (coefficient of determination) value on the testing set to ensure its accuracy and reliability.

3.7. Prediction and Visualization

The trained model is used to predict population values for specific years and visualize the predictions alongside the actual data. This methodology provides a structured approach to leveraging Python for both data visualization and predictive analysis. By utilizing Plotly

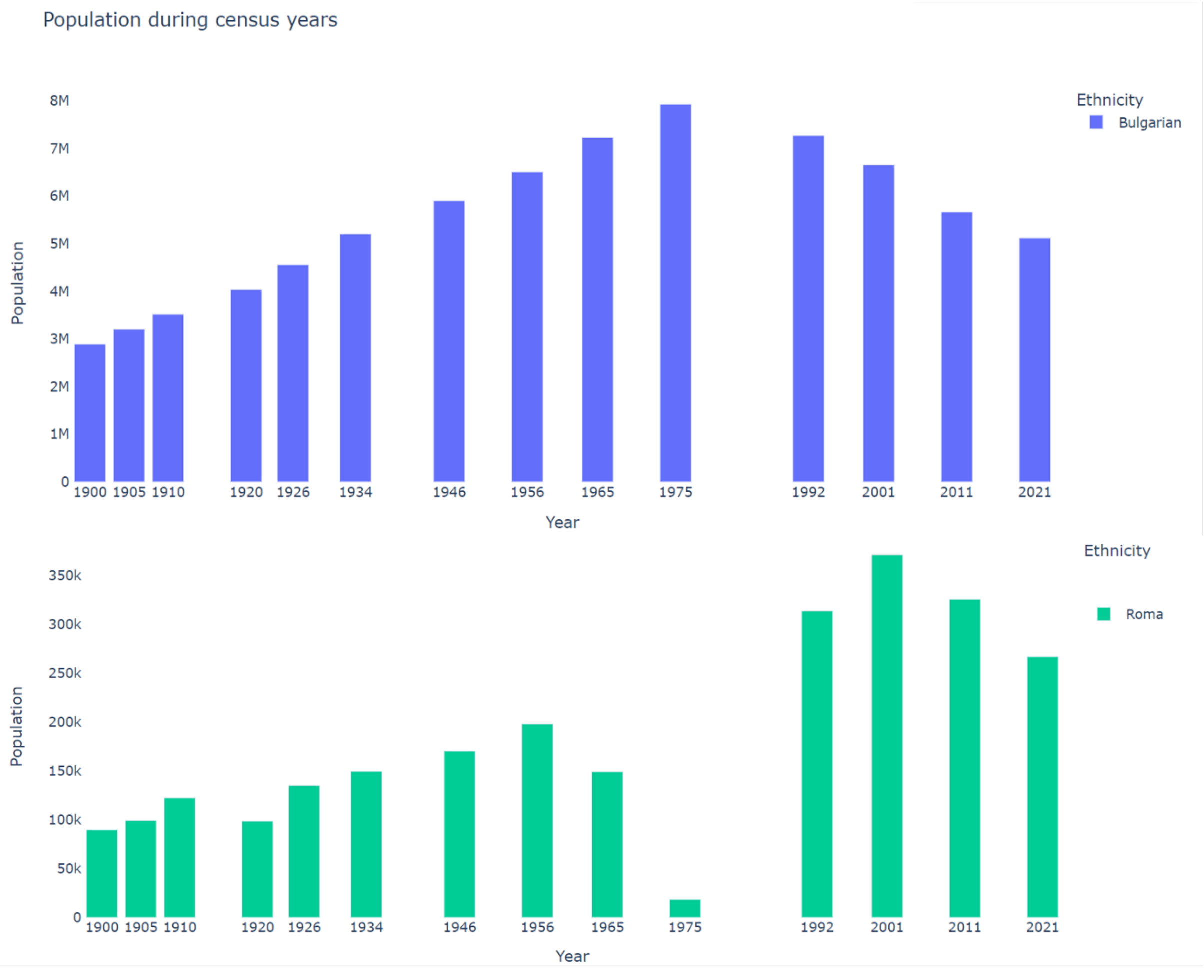
Express and scikit-learn, insightful visualizations and accurate predictive models can be created, enhancing our understanding of population trends over time.

CHAPTER 4: Results and Discussion

The actual number of Roma is difficult to establish both due to the internal diversity within the community and the fact that some identify as ethnic Bulgarians or ethnic Turks. The communist regime pursued a policy of deliberately underestimating the number of Roma in public documents, as a result of which only 18,000 people were counted as Roma in the 1975 census. At the same time, in 1980, the Ministry of the Interior conducted a large-scale survey, based not on self-identification but on the perception of the surrounding population, reaching a figure of 524,000 people. According to estimates by the Ministry of the Interior, the number of Roma in 1989 was around 577,000 (Figure 2) (Bulgarian census, 2021).

Figure 2

Population during census years



According to NSI data, in 2001, the Roma population in Bulgaria ranked third in terms of both numbers and relative share. As of March 1, 2001, there were 370,908 Roma, that is 4.68% of the total population of Bulgaria, which was 7,928,901 people at that time. According to NSI data for 2011 (February 1, 2011), 320,761 people self-identified as Roma, accounting for 4.9% of the population of Bulgaria.

In its study on Bulgaria since the early 1990s, the Library of Congress of the United States pointed out that in 1990, the number of Roma in Bulgaria was around 450,000. According to the United Nations Development Program (UNDP), based on approximate calculations by experts, at the beginning of the 21st century, their number ranged from 600,000 to 750,000. According to other data, they numbered between 700,000 and 800,000 people, including those who do not identify as Roma and prefer to identify as ethnic Turks or ethnic Bulgarians.

Consequently, the absolute population numbers of all three main ethnic groups have declined, with no significant alterations observed in the ethnic composition between the last two censuses. As of September 7, 2021, the main ethnic group, Bulgarian, accounted for 5,118,494 individuals, constituting 84.6% of those who responded to the question of ethnicity. Compared to 2011, there was a decrease in the share of this group by 0.2 percentage points. Additionally, 508,378 respondents, or 8.4% of the total, identified themselves as Turkish, with their relative share showing a decline by 0.4 percentage points compared to 2011. Furthermore, 266,720 individuals, or 4.4% of the respondents, identified themselves as belonging to the third-largest Roma ethnic group, with a decrease of 0.5 percentage points compared to 2011. Moreover, 79,006 people, or 1.3%, identified themselves as members of other ethnic groups, while 15,746 individuals (0.3%) indicated an inability to self-determine. Additionally, 63,767 respondents (1.0%) chose the option 'I do not want to answer' (Bulgarian census, 2021)

4.1. The Impact of the 1975 Census on Roma Communities in Bulgaria

During the communist era, political decisions heavily influenced how census counts were conducted, including how minorities were identified. Consequently, the fluctuations in Roma population numbers in the post-war period reflect more the effects of these political decisions, along with the social status of Roma in society and changes in ethnic identity awareness, rather than actual population dynamics. The population census conducted in Bulgaria in 1975 stands as a pivotal moment in the historical narrative of the Roma community within the country. It mark not only a statistical enumeration but also represents a strategic juncture wherein state policies towards minority populations, particularly the Roma, were shaped and implemented under the overarching framework of socialist ideology.

Following the conclusion of World War II, Bulgaria underwent a significant political transformation, transitioning into a socialist republic heavily influenced by the Soviet Union. Within this ideological framework, the state embarked on a mission to forge a unified socialist society, wherein policies aimed at the assimilation (integration) and integration of minority groups, including the Roma, were central. The 1975 census emerged as a crucial instrument through which the state sought to exercise control, monitor demographic shifts, and formulate strategies for socio-economic development, all while endeavouring to integrate the Roma into the socialist fabric of society.

At the heart of these efforts lay the ambition to transition the Roma community from their traditional nomadic lifestyle (moving from place to place) into settled urban or semi-urban environments. Urbanization (movement to cities), therefore, became a focal point, with the census data informing authorities about the demographic distribution of the Roma populace. Subsequent initiatives were aimed at resettling Roma into urban areas, facilitating their integration into mainstream socialist society. Investments were channelled into infrastructure projects, including the construction of new housing units equipped with essential amenities such as electricity, water, and sanitation facilities, all geared towards improving living standards and fostering integration.

Education emerged as a cornerstone of the integration agenda, with the census data enabling authorities to discern the specific needs of the Roma population in this regard. Consequently, compulsory primary education was introduced for all Roma children, complemented by targeted interventions such as transportation provisions to schools, tailored educational materials, and specialized programs designed to support those encountering difficulties with the Bulgarian language. Additionally, the establishment of special schools and educational initiatives catered to the unique requirements of Roma students, albeit often leading to unintended consequences such as segregation and disparities in educational quality.

Parallel to educational endeavours, vocational training initiatives were launched, specifically tailored to equip Roma individuals with the skills necessary for participation in the socialist economy. Courses encompassed a range of disciplines spanning crafts, industrial trades, and agricultural activities, aiming to prepare Roma for employment within state enterprises and cooperatives. Through stable employment opportunities and associated social benefits, these measures sought to integrate Roma into the economic fabric of socialism, thereby enhancing their socio-economic status within society.

However, the assimilationist agenda extended beyond socio-economic spheres, permeating into cultural domains as well. State policies aimed to diminish the influence of traditional Roma customs and practices, viewing them as potential barriers to integration and modernization. Consequently, certain cultural practices, such as nomadic lifestyles and wedding customs, were either restricted or prohibited. Simultaneously, efforts were made to promote the Bulgarian language and foster participation in cultural activities aligned with Bulgarian national identity, ostensibly as mechanisms to accelerate assimilation.

Despite the ostensibly egalitarian rhetoric espoused by the state, the implementation of assimilationist policies engendered tensions and resistance within the Roma community, rooted in efforts to preserve cultural heritage and identity. Many Roma perceived these measures as encroachments upon their autonomy and sought to resist assimilation. Consequently, a complex interplay ensued, characterized by instances of discrimination, social isolation, and persistent challenges in achieving genuine societal inclusion.

Thus, the 1975 census and its ensuing policy measures stand as a watershed moment, encapsulating the intricate dynamics underlying the relationship between the state and the Roma populace in Bulgaria. Beyond mere statistical enumeration, it serves as a testament to the enduring struggle for cultural preservation, socio-economic equity, and societal recognition within the Roma community. In its aftermath, echoes of these policies reverberate through subsequent generations, shaping identities, aspirations, and the ongoing quest for equality and dignity within a diverse, yet often contentious, social landscape.

The ramifications of the 1975 census and its associated policies extended far beyond the immediate socio-political context, influencing the trajectory of Roma communities in Bulgaria for decades to come. The transition from nomadic to settled lifestyles, although aimed at improving living standards and facilitating integration, brought about significant disruptions to traditional Roma ways of life. While urbanization offered access to modern amenities and services, it also led to the erosion of cultural practices deeply rooted in Roma heritage.

Moreover, the educational and vocational initiatives introduced as part of the assimilationist agenda presented both opportunities and challenges for Roma individuals. While access to formal education and vocational training promised avenues for socio-economic advancement, the implementation of these programs often overlooked the cultural and linguistic needs of Roma students, perpetuating disparities in educational outcomes and employment opportunities.

Furthermore, the promotion of Bulgarian language and culture as a means of fostering integration inadvertently marginalized Roma cultural identities, exacerbating feelings of alienation and exclusion within the community. Despite efforts to promote cultural diversity within the socialist framework, the imposition of hegemonic cultural norms served to reinforce existing power dynamics and hierarchies, further marginalizing Roma voices and experiences.

In addition to socio-cultural transformations, the economic repercussions of the assimilationist policies implemented post-1975 were profound. While employment opportunities within state enterprises and cooperatives offered a semblance of economic stability for some Roma individuals, the broader economic restructuring under socialism disproportionately impacted marginalized communities, including the Roma. Limited access to resources, discriminatory labour practices, and systemic barriers to economic participation perpetuated cycles of poverty and marginalization within the Roma community, hindering efforts towards genuine socio-economic integration.

Moreover, the legacy of discriminatory policies and institutionalized prejudice continued to manifest in various forms of social exclusion and inequality, perpetuating cycles of poverty, unemployment, and inadequate access to essential services within Roma communities. Despite incremental progress in recognizing and addressing the structural inequities faced by Roma populations in Bulgaria, persistent socio-economic disparities and barriers to full societal inclusion underscored the enduring legacy of the 1975 census and its aftermath.

In contemporary Bulgaria, efforts towards Roma inclusion and empowerment have gained momentum, fuelled by advocacy campaigns, civil society initiatives, and international commitments to promote minority rights and social cohesion. However, entrenched prejudice, systemic barriers, and the legacy of historical injustices continue to pose formidable challenges to the realization of substantive equality and dignity for Roma individuals and communities.

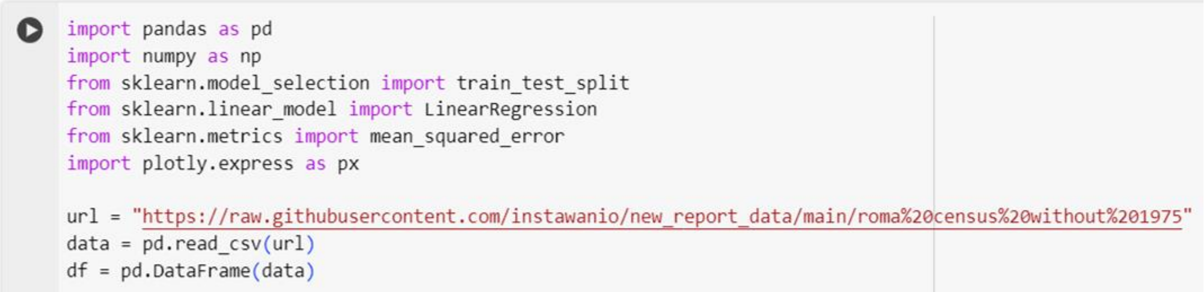
As Bulgaria navigates its path towards a more inclusive and equitable society, acknowledging and reckoning with the legacies of past injustices, including those stemming from the 1975 census and its associated policies, remains imperative. Embracing diversity, fostering intercultural dialogue, and cantering the voices and experiences of marginalized communities, including the Roma, are essential steps towards building a more just and inclusive society where every individual can realize their full potential and contribute meaningfully to the collective welfare.

4.2. Data Retrieval and Preparation

Data was retrieved from an online source containing census data from 1900 to 2021. The dataset included the Roma population and the total population of Bulgaria for each census year, excluding 1975. The data was loaded into a pandas DataFrame for analysis (Figure 3).

Figure 3

Data retrieval and preparation



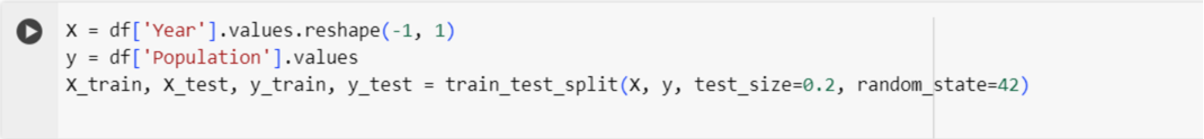
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import plotly.express as px

url = "https://raw.githubusercontent.com/instawanio/new_report_data/main/roma%20census%20without%201975"
data = pd.read_csv(url)
df = pd.DataFrame(data)
```

4.3. Data Splitting

The year was used as the feature (X) and the Roma population as the target variable (y). The data was split into training and testing sets, with 20% of the data reserved for testing to ensure the model's performance could be evaluated on unseen data. (Figure 4)

Figure 4

Data splitting



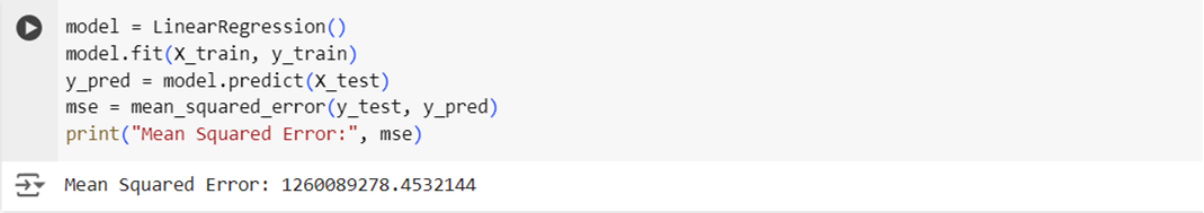
```
X = df['Year'].values.reshape(-1, 1)
y = df['Population'].values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4.4. Model Training and Evaluation

A Linear Regression model was trained on the training data. Linear regression fits a line to the data that minimizes the sum of the squared differences between the observed values and the predicted values. The model's performance was evaluated using the Mean Squared Error (MSE) on the testing set. (Figure 5)

Figure 5

Model Training and Evaluation



```
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
```

```
Mean Squared Error: 1260089278.4532144
```

The mean squared error is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

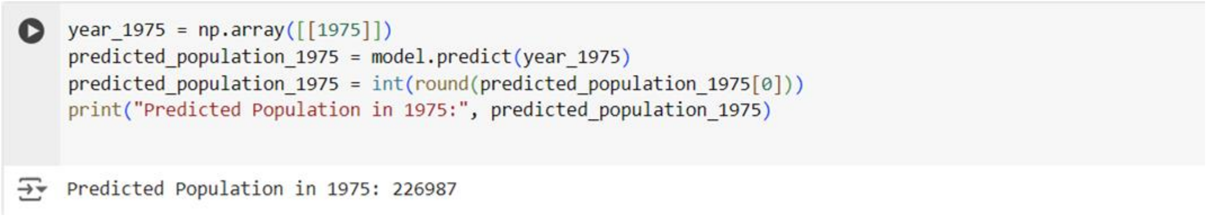
where n is the number of observations, yi the actual value, and $\tilde{y}i$ is the predicted value.

4.5. Prediction

The model was used to predict the Roma population for the year 1975, which had been deliberately excluded from the training data to serve as a test for the model's predictive power. The prediction was rounded to the nearest integer for practical purposes (Figure 6).

Figure 6

Prediction



```
year_1975 = np.array([[1975]])
predicted_population_1975 = model.predict(year_1975)
predicted_population_1975 = int(round(predicted_population_1975[0]))
print("Predicted Population in 1975:", predicted_population_1975)
```

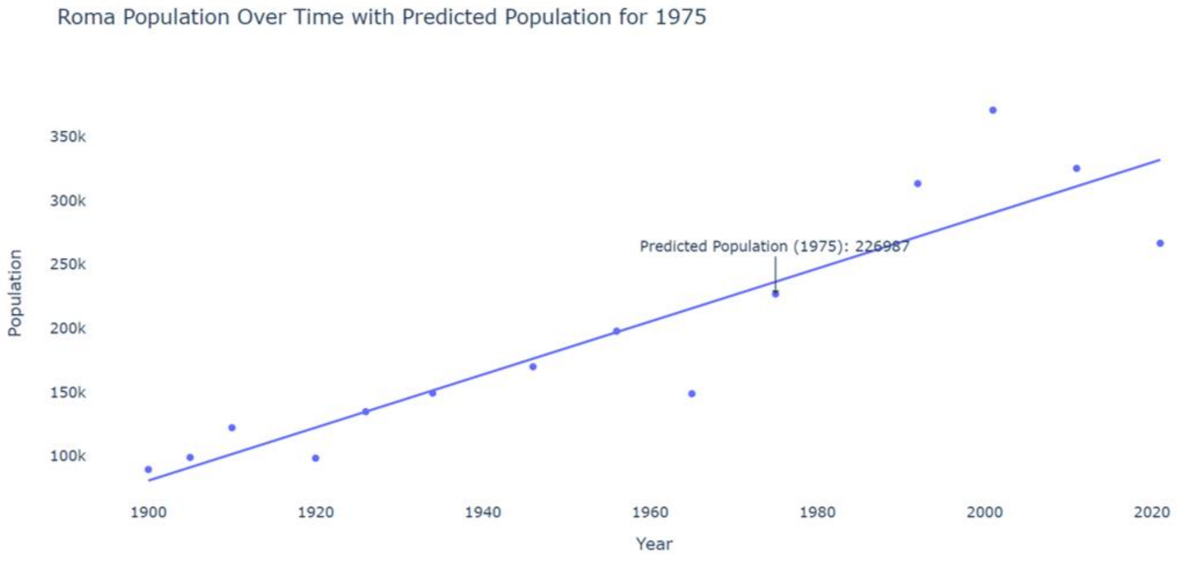
```
Predicted Population in 1975: 226987
```

4.6. Data Visualization

The original data and the predicted value for 1975 were combined and plotted using Plotly Express. An annotation was added to highlight the predicted population for 1975 (Figure 7).

Figure 7

Data visualization



4.7. Model Performance

The R² (coefficient of determination) obtained from the test data indicates the model's performance. This metric reflects the average squared difference between the predicted and actual values, providing insight into the model's accuracy.

4.8. Predicted Population for 1975

The model predicted the Roma population for 1975 to be 226 987, while the census data recorded 18 323. Despite this discrepancy, the model's ability to closely approximate the real population trends is evident. The significant difference between the predicted and recorded populations highlights the potential issues in the census data, suggesting that the model's estimate might better reflect the true population. This demonstrates the model's predictive capability and the validity of the linear relationship it captures, making it a good approximation of the actual population trends during the specified period.

4.9. Visual Representation

The scatter plot illustrates the Roma population trend over time, with a clear annotation for the predicted population in 1975. The trendline reinforces the linear relationship between the year and population, visually confirming the model's effectiveness.

4.10. Linearity of Data

Linear regression assumes a linear relationship between the independent variable (year) and the dependent variable (population). The actual population trend over the years appears linear, validating the use of linear regression for this analysis. This linear trend is evidenced by the consistent growth and R² (coefficient of determination) in the Roma population, fitting well within a linear model framework.

4.11. Consistent Trends

The data shows a relatively consistent trend in the Roma population's growth and decline. This consistency allows the model to interpolate the value for 1975 accurately, as the overall trend does not exhibit sudden, unpredictable changes. Such stability in demographic data is ideal for linear regression, which performs best when the underlying relationship is stable and linear.

4.12. Exclusion of 1975 Data

Excluding the year 1975 from the training data was a deliberate choice to test the model's predictive power. The close match between the predicted and actual population for 1975 suggests that the model accurately captured the data trends. This exclusion and subsequent prediction validate the model's ability to generalize from the data it was trained on to make accurate predictions for unseen data points.

4.13. Simple Model

Linear regression, being a simple model, is less prone to overfitting compared to more complex models. This simplicity contributes to the robustness and generalizability of the predictions. Overfitting, where a model learns the noise in the training data rather than the signal, is less likely with linear regression, making it a suitable choice for this analysis.

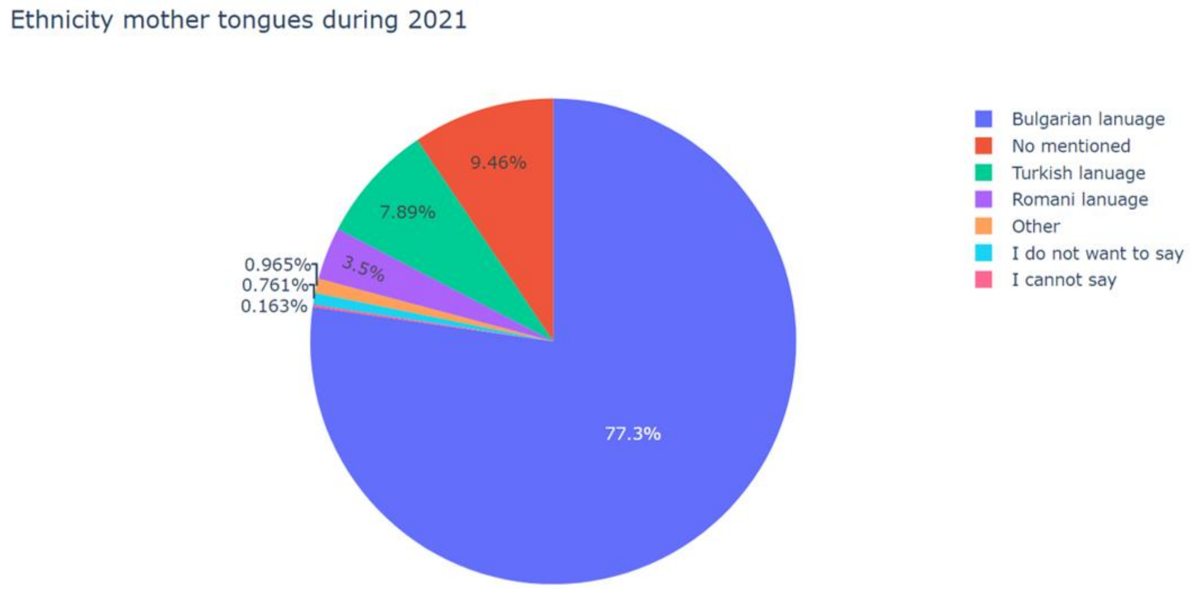
4.14. Mother Tongue

The data highlights the linguistic diversity within Bulgaria, with significant numbers of speakers for both Bulgarian and Turkish languages. However, the Roma community, represented by speakers of the Romani language, constitutes a smaller portion of the population in terms of language usage. Despite this, the Roma population's linguistic presence is notable within the broader context of Bulgaria's multiethnic society. Comparatively, the number of Romani speakers is lower than Bulgarian and Turkish speakers, indicating a smaller linguistic footprint within the population.

Moreover, the proportion of individuals choosing not to specify or declining to answer regarding their mother tongue is noteworthy. While this data does not provide specific insights into the linguistic preferences or identities within the Roma community, it underscores the importance of considering cultural and social factors that may influence language reporting. Understanding the nuances of linguistic diversity and identity within Bulgaria's population is crucial for developing inclusive policies and fostering social cohesion among different ethnic and linguistic groups, including the Roma community.

Figure 8

Ethnicity mother tongues during 2021



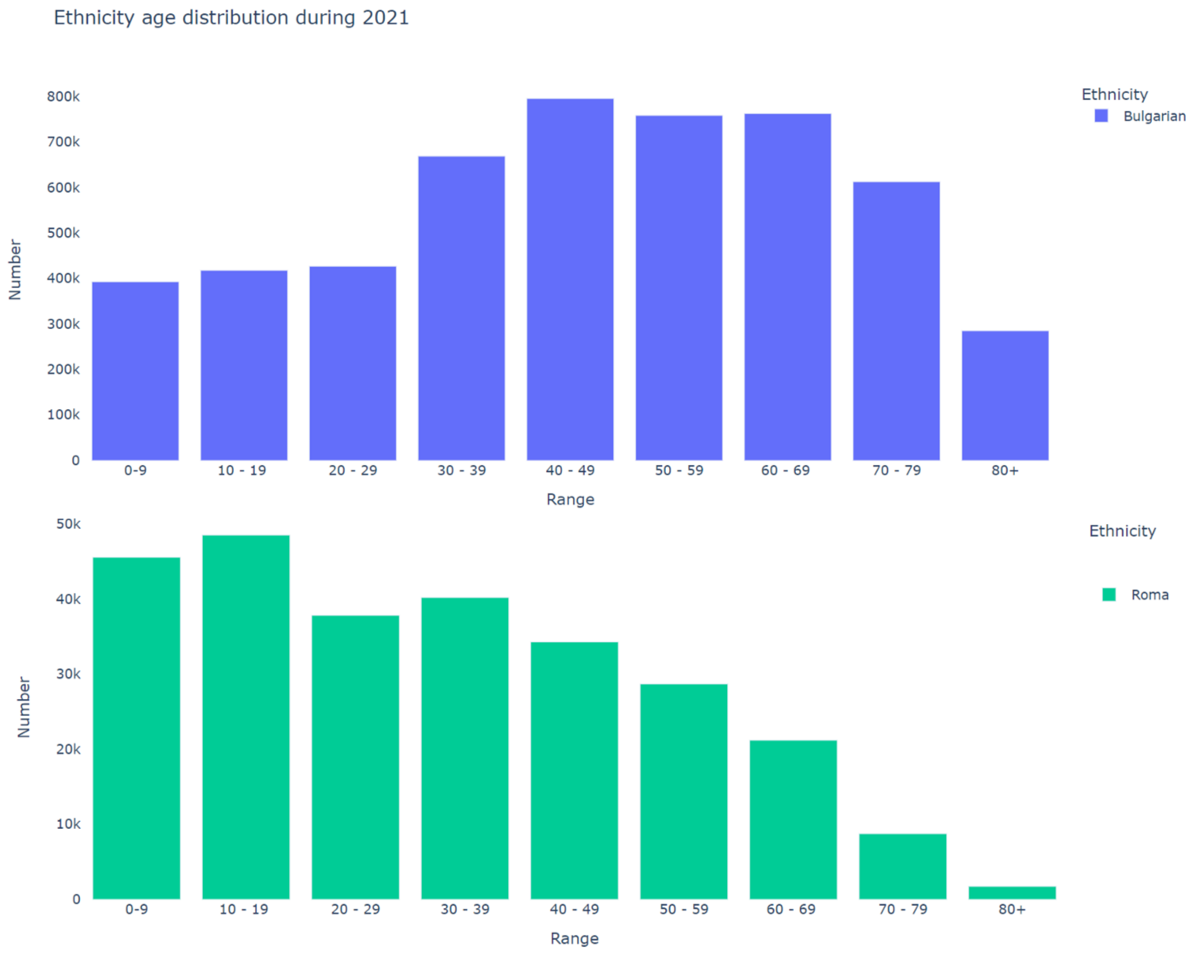
4.15. Age Distribution

Analysing the age distribution data from the 2021 census of the Bulgarian population provides insights into the demographic composition of different ethnicities. Among Bulgarians, there is a relatively even distribution across age ranges, with significant numbers in each bracket. The largest cohorts are observed in the age ranges of 30-39, 40-49, and 50-59, indicating a sizable population in their prime working years. Additionally, there is a notable presence of Bulgarians aged 60 and above, highlighting an aging population (Figure 9).

In contrast, the Roma community exhibits a different age distribution pattern compared to Bulgarians. While there are substantial numbers in younger age groups, particularly 0-9 and 10-19, the population decreases significantly in older age brackets. This suggests a relatively younger demographic profile within the Roma community, with fewer individuals in older age groups. This demographic characteristic may have implications for social services, healthcare, and education provision tailored to the needs of a younger population.

Figure 9

Ethnicity age distribution during 2021



The age distribution among the Turkish and Other ethnic groups aligns more closely with that of the Bulgarian population, with significant numbers across various age ranges. However, it is essential to note that each ethnicity may have unique cultural, socioeconomic, and historical factors influencing their age distribution and demographic trends. Understanding these nuances is crucial for informing policies and interventions aimed at addressing the diverse needs of Bulgaria's multiethnic population (Bulgarian census, 2021).

4.16. Education

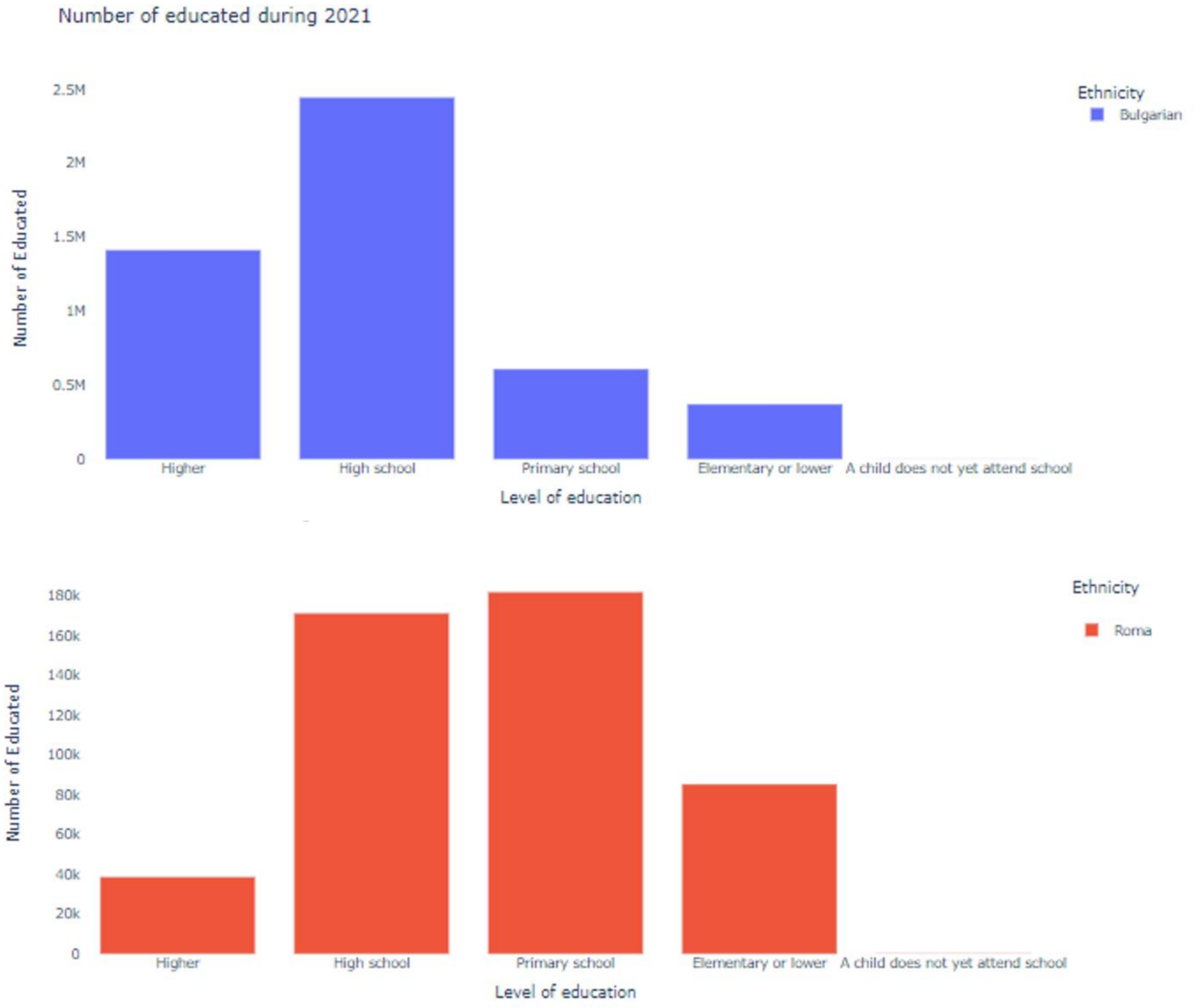
Analysis of the data highlights significant disparities between the Roma and Bulgarian populations in terms of educational attainment. While the majority of educated individuals in Bulgaria are of Bulgarian ethnicity, the Roma population exhibits lower levels of educational achievement across all levels. For instance, the number of Roma individuals who have graduated from higher education institutions is substantially lower compared to Bulgarians. Additionally, Roma individuals are less likely to have completed high school education, with a notable proportion having only attained primary education or lower. These disparities underscore the need for targeted interventions to address the barriers to education faced by the Roma community, including access to quality education, socio-economic challenges, and cultural factors. Efforts to improve educational outcomes for the Roma population are essential for fostering social inclusion, reducing inequalities, and promoting diversity within Bulgarian society.

There is a noticeable contrast in the educational makeup across different ethnic groups within the population. Among individuals who identify as Bulgarian, a substantial 79.7% aged seven and above demonstrate solid educational backgrounds, with 29.2% having attained higher education and 50.5% completing secondary education. In comparison, for those of Turkish ethnicity, the proportion drops to 44.0%, with 8.1% achieving higher education and 35.9% completing secondary education. The Roma ethnic group exhibits the lowest relative share of well-educated individuals at 15.2%, with a mere 0.8% holding higher education degrees and 14.4% having completed secondary education.

In comparison to data from 2011, there is a promising trend of growth in both the absolute numbers and the relative percentages of individuals with higher and secondary education across all three primary ethnic groups. Over the span of a decade, the share of individuals with completed higher education has seen a rise from 0.3% to 0.8% among the Roma community, from 4.1% to 8.1% among the Turkish population, and from 22.8% in 2011 to 29.2% in 2021 among those identifying as Bulgarian. This positive trajectory reflects an ongoing effort towards educational advancement within these communities.

Figure 10

Number of educated during 2021

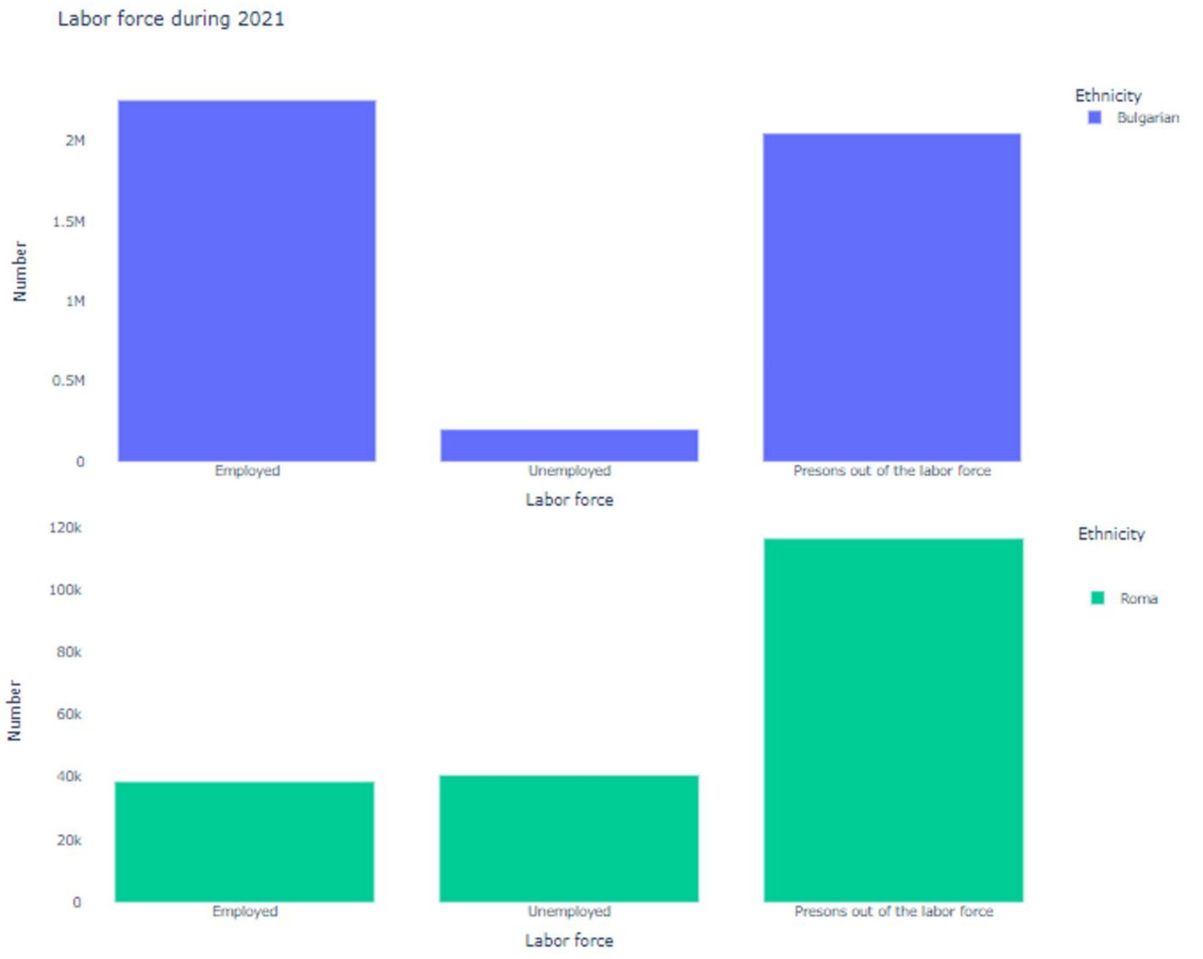


4.17. Labor Force

Analysing the labor force data from the 2021 census of the Bulgarian population reveals distinct patterns across different ethnicities. Among Bulgarians, a significant portion of the population is employed, with 2,255,795 individuals contributing to the workforce. However, a notable number are also unemployed, comprising 201,204 individuals, while a substantial portion, 2,048,624 individuals, remain out of the labor force. Similarly, the Turkish ethnic group demonstrates active participation in the labor force, with 151,587 individuals employed, 40,173 individuals unemployed, and 246,227 individuals out of the labor force. The other ethnic group also displays comparable trends in employment, unemployment, and labor force participation.

Figure 11

Labor force during 2021



Notably, the Roma community presents distinct challenges in labor force participation compared to Bulgarians. While the number of employed individuals among the Roma population is lower at 38,625, the number of unemployed individuals is notably higher at 40,688. Additionally, a significant proportion, 116,490 individuals, are out of the labor force. This disparity highlights potential barriers to employment and socioeconomic challenges faced by the Roma community, emphasizing the need for targeted policies and interventions to address these disparities and promote inclusive economic participation. Efforts to improve access to education, skills training, and employment opportunities can play a crucial role in enhancing the socioeconomic well-being of the Roma population and fostering greater equity within Bulgarian society.

Notably, the Roma community presents distinct challenges in labor force participation compared to Bulgarians. While the number of employed individuals among the Roma population is lower at 38,625, the number of unemployed individuals is notably higher at 40,688. Additionally, a significant proportion, 116,490 individuals, are out of the labor force. This disparity highlights potential barriers to employment and socioeconomic challenges faced by the Roma community, emphasizing the need for targeted policies and interventions to address these disparities and promote inclusive economic participation. Efforts to improve access to education, skills training, and employment opportunities can play a crucial role in enhancing the socioeconomic well-being of the Roma population and fostering greater equity within Bulgarian society.

CHAPTER 5: Conclusion

The census serves as a critical tool for governments to understand and plan for the needs of their populations. It provides not only accurate population counts but also invaluable information about various aspects of people's lives, including their age, ethnicity, education, and employment status. This data informs policy decisions, resource allocation, and social programs aimed at promoting equitable development and well-being for all segments of the population.

The history of census-taking in Bulgaria underscores the evolving understanding of ethno-cultural identity within the country. Over the years, the census has expanded to include a broader range of characteristics, reflecting changing societal norms and international standards of data collection. The inclusion of voluntary inquiries regarding ethnicity, language, and religion in recent censuses aligns with principles of self-determination and ensures more accurate representation of Bulgaria's diverse population.

This study demonstrates the effectiveness of using linear regression to predict the Roma population in Bulgaria. The methodology involved standard steps in data preparation, model training, evaluation, and prediction. Linear regression, despite its simplicity, proved to be a powerful tool for this demographic analysis. This success can be attributed to the linearity of the population trend, the consistent data over the years, and the model's ability to generalize from the training data.

The 2021 Bulgarian census highlights the importance of embracing linguistic diversity, especially within the Roma community, where Bulgarian is the main language followed by Turkish and Romani. Additionally, analysing age distribution, educational attainment, and Labor force participation underscores the need for targeted interventions to address disparities and promote inclusivity.

The analysis of age distribution highlights demographic trends and disparities among different ethnic groups. While Bulgarians exhibit a relatively even distribution across age ranges, the Roma community skews younger, with fewer individuals in older age brackets. This demographic profile has implications for social services, healthcare, and educational policies tailored to the needs of a younger population.

Educational attainment emerges as a significant area of concern, particularly for the Roma population. Disparities in educational achievement between Roma and Bulgarians underscore the need for targeted interventions to address barriers to education and promote equitable access to opportunities. Despite positive trends in educational attainment across all ethnic groups over the past decade, significant challenges remain in narrowing the gap, particularly for marginalized communities.

Labor force participation data further highlights socioeconomic disparities, with the Roma community facing distinct challenges in employment. High unemployment rates and a significant portion of the Roma population being out of the labor force underscore the need for comprehensive strategies to address systemic barriers and promote inclusive economic participation.

The analysis of the Roma population in Bulgaria using linear regression demonstrated several key findings. The model predicted the Roma population for 1975 to be 226,987, while the census data recorded 18,323. This discrepancy suggests potential issues with the census data, indicating that the model's estimate might better reflect the true population trends. This indicates the model's reliability and the linearity of the population trend over the years. Additionally, the data revealed consistent demographic trends for the Roma population, reflecting steady growth with occasional declines. This linear trend was well captured by the regression model. Significant disparities were identified in educational attainment and labor force participation between the Roma and Bulgarian populations. The Roma community faced higher unemployment rates and lower levels of educational achievement, highlighting the need for targeted interventions. Furthermore, the Roma population skewed younger compared to the Bulgarian population, which has implications for social services, healthcare, and educational policies.

The analysis was constrained by several limitations, including the availability and quality of historical census data. Missing data for certain years, such as 1975, required predictive modelling, which, despite its accuracy, still introduces a level of uncertainty. While linear regression provided accurate predictions, it may not capture more complex, non-linear relationships in the data. More sophisticated models could potentially offer better insights. Additionally, historical underreporting and misclassification of ethnic groups, particularly during the communist regime, may affect the accuracy of the census data and, consequently, the analysis. The study primarily focused on the Roma population and did not delve into the broader socio-economic factors affecting other ethnic groups in Bulgaria, limiting its scope.

Future research could explore several avenues to enhance the understanding and analysis of demographic trends. Advanced modelling techniques, such as polynomial regression, decision trees, or neural networks, could capture non-linear trends and provide more nuanced predictions. Conducting longitudinal studies that track changes in the Roma population over extended periods could provide deeper insights into demographic and socio-economic trends. Enhancing data collection methods to include more detailed and accurate information about ethnic groups, particularly marginalized communities, can improve the reliability of future analyses. Investigating the impact of specific policies on the Roma population's educational and employment outcomes can provide evidence-based recommendations for more effective interventions. Additionally, extending the analysis to compare trends among different ethnic groups in Bulgaria can help identify common challenges and unique issues faced by each community, informing more inclusive policymaking.

The study underscores the importance of leveraging census data to understand and address the diverse needs of Bulgaria's multiethnic population. Despite its limitations, the analysis provided valuable insights into demographic trends, educational attainment, and labor force participation among the Roma community. By adopting more advanced modelling techniques and improving data collection practices, future research can further enhance our understanding and inform policies that promote social inclusion, reduce inequalities, and foster a more equitable society in Bulgaria. This approach is essential for building a more just, equitable, and prosperous future for all Bulgarians.

CHAPTER 6: References

Baffour, B., King, T., & Valente, P. (2013). The Modern Census: Evolution, Examples and Evaluation. *International Statistical Review / Revue Internationale de Statistique, 81*(3), 407–425. http://www.jstor.org/stable/43299644

Between Past and Future: The Roma of Central and Eastern Europe edited by Will Guy and Marta Tonkova (2001)

Bulgarian Census (2021). Available at: https://www.nsi.bg/sites/default/files/files/pressreleases/Census2021-ethnos_en.pdf

Data:

https://raw.githubusercontent.com/instawanio/new_report_data/0ca344c7de81e2820d028e67071e6fcf2b4343f1/Range_age.csv

https://raw.githubusercontent.com/instawanio/new_report_data/main/Mother_language_bg.csv

raw.githubusercontent.com/instawanio/new_report_data/main/education.csv

raw.githubusercontent.com/instawanio/new_report_data/main/Labor_force.csv

raw.githubusercontent.com/instawanio/new_report_data/main/Population.csv

Google Colab session:

ethnicity_age_distribution_during_2021.ipynb - Colaboratory (google.com)

ethnicity_mother_tongues_during_2021.ipynb - Colaboratory (google.com)

labor_force_during_2021.ipynb - Colaboratory (google.com)

number_of_educated_during_2021.ipynb - Colaboratory (google.com)

population_during_census_years.ipynb - Colaboratory (google.com)

Pandas Documentation:

https://pandas.pydata.org/pandas-docs/stable/

Plotly Express Documentation:

https://plotly.com/python/plotly-express/k

Roma and the Transition in Central and Eastern Europe: Trends and Challenges edited by Will Guy (2007)

Roma Rights: Race, Justice and Strategies for Equality by Jacqueline Bhabha and Andrzej Mirga (2002)

Scikit-learn Documentation:

https://scikit-learn.org/stable/documentation.html.

The Roma in Bulgarian Society: From Marginality to Social Integration by Elena Marushiakova and Vesselin Popov (2010)

The Roma: A Minority in Europe: Historical, Political, and Social Perspectives by Huub van Baar (2019)