DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Census 2011 Population Report

Submitted by

Trijal Luhariwala
Registration No 12313979
Programme P132: B. Tech
Section K23KM
Course Code INT375

Under the Guidance of

Anchal Kaundal (UID: 29612)

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Trijal Luhariwala bearing Registration no. 12313979 has completed INT375: Data Science Toolbox: Python Programming project titled, “***Census 2011 Population Report***” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/his original development, effort and study.

Anchal Kaundal
Assistant Professor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 12-04-2025

DECLARATION

I, Trijal Luhariwala, student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 Signature

Registration No. 12313979 Trijal Luhariwala

Acknowledgement

I would like to express my sincere gratitude to Lovely Professional University for providing the opportunity and platform to undertake this project.

I am especially thankful to Anchal Kaundal for his invaluable guidance, constant support, and insightful feedback throughout the course of this project. His mentorship played a crucial role in shaping the direction and success of my work. I also wish to thank my faculty mentors, peers, and friends for their encouragement and helpful suggestions during the project.

This experience has significantly enhanced my technical, analytical, and visualization skills, particularly in working with Excel dashboards. Finally, I extend my heartfelt thanks to everyone who directly or indirectly contributed to the successful completion of this project.

Introduction

The Census of India is one of the most comprehensive demographic exercises in the world, capturing the social, economic, and cultural contours of a rapidly evolving nation. The 2011 census, the 15th in India's history, provides a detailed snapshot of over 1.2 billion people, making it an invaluable resource for policymakers, researchers, and analysts. Beyond mere population counts, this dataset reveals critical insights into literacy rates, workforce distribution, gender disparities, and regional development patterns—each of which plays a crucial role in shaping India's growth trajectory.

This project leverages Python's data analysis capabilities to uncover meaningful trends within the 2011 census data. By employing libraries like Pandas for data manipulation and Matplotlib/Seaborn for visualization, we transform raw numbers into actionable insights. The goal is not just to observe statistical patterns but to understand their real-world implications—whether it's the urban-rural divide in education, the uneven distribution of employment opportunities, or the societal factors influencing women's workforce participation.

Demographic data, when analyzed thoughtfully, can reveal stories that numbers alone cannot tell. For instance, how does literacy correlate with employment in different regions? Are states with higher female literacy rates also those where more women join the workforce? By exploring these questions, this project bridges the gap between data and human impact, offering a clearer picture of India's developmental challenges and opportunities.

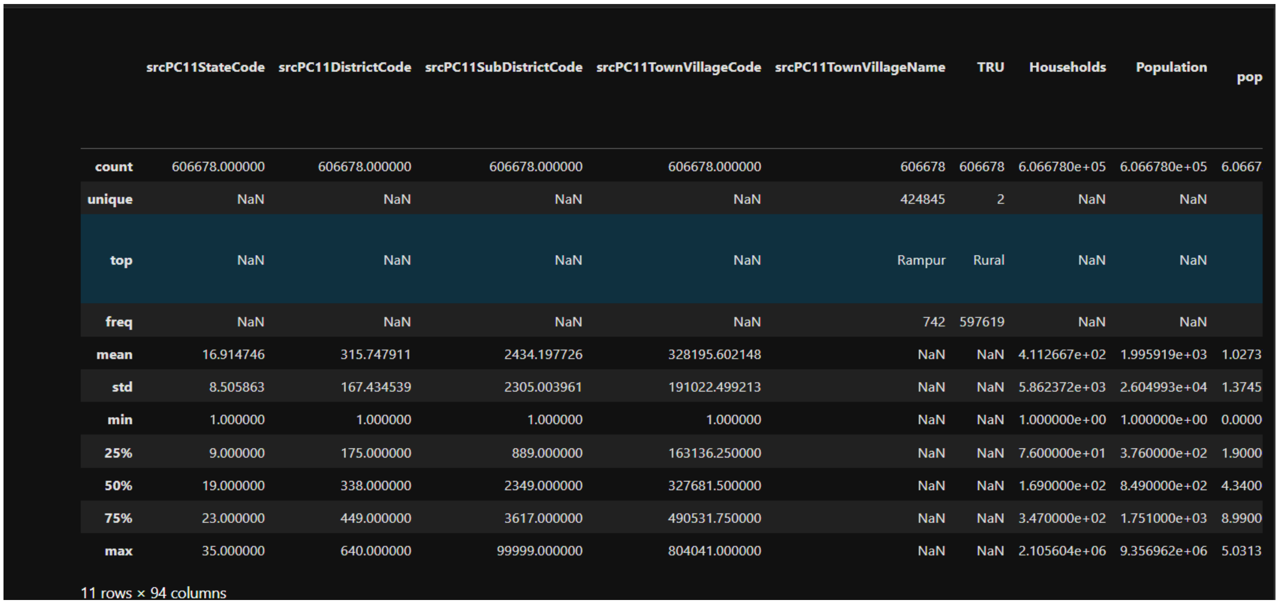
The analysis focuses on key themes: population structure, education, employment, and gender dynamics. Rather than just presenting charts and figures, we aim to contextualize the findings—highlighting why certain trends matter and what they signify for future policy decisions. Whether you're a researcher, student, or policymaker, this exploration provides a foundation for deeper inquiry into India's demographic evolution.

Ultimately, this project is as much about understanding the past as it is about planning for the future. The 2011 census data, though over a decade old, still holds relevant lessons for contemporary discussions on equitable development. By applying modern analytical tools to this rich dataset, we can uncover patterns that help inform smarter, data-driven decisions for years to come.

Source of Data

The dataset used in this analysis comes from the **National Data & Analytics Platform (NDAP)**, an initiative by NITI Aayog, Government of India, to provide accessible and standardized datasets for research and policymaking. The specific dataset, titled **"Census 2011 - Town and Village Level Data,"** can be accessed at: https://ndap.niti.gov.in/dataset/6000

This dataset provides granular demographic, social, and economic data at the town and village level, making it invaluable for micro-level analysis. It includes key metrics such as population distribution, literacy rates, workforce participation, and caste/tribe demographics. The data is sourced directly from the **Registrar General and Census Commissioner of India**, ensuring its authenticity and reliability for research.



	srcPC11StateCode	srcPC11DistrictCode	srcPC11SubDistrictCode	srcPC11TownVillageCode	srcPC11TownVillageName	TRU	Households	Population	pop
count	606678.000000	606678.000000	606678.000000	606678.000000	606678	606678	6.066780e+05	6.066780e+05	6.0667
unique	NaN	NaN	NaN	NaN	424845	2	NaN	NaN	
top	NaN	NaN	NaN	NaN	Rampur	Rural	NaN	NaN	
freq	NaN	NaN	NaN	NaN	742	597619	NaN	NaN	
mean	16.914746	315.747911	2434.197726	328195.602148	NaN	NaN	4.112667e+02	1.995919e+03	1.0273
std	8.505863	167.434539	2305.003961	191022.499213	NaN	NaN	5.862372e+03	2.604993e+04	1.3745
min	1.000000	1.000000	1.000000	1.000000	NaN	NaN	1.000000e+00	1.000000e+00	0.0000
25%	9.000000	175.000000	889.000000	163136.250000	NaN	NaN	7.600000e+01	3.760000e+02	1.9000
50%	19.000000	338.000000	2349.000000	327681.500000	NaN	NaN	1.690000e+02	8.490000e+02	4.3400
75%	23.000000	449.000000	3617.000000	490531.750000	NaN	NaN	3.470000e+02	1.751000e+03	8.9900
max	35.000000	640.000000	99999.000000	804041.000000	NaN	NaN	2.105604e+06	9.356962e+06	5.0313

11 rows × 94 columns

The dataset contains **94 columns**, broadly categorized into:

1. Geographic Identifiers

- srcPC11StateCode, srcPC11DistrictCode, srcPC11SubDistrictCode, srcPC11TownVillageCode: Unique codes for state, district, sub-district, and town/village.
- srcPC11TownVillageName: Name of the town/village.
- TRU: Classification as **Rural (R), or Urban (U)**.
-

2. Population Demographics

- Households: Total number of households.
- Population: Total population count.

- Male population, Female population: Gender-wise distribution.
- Population in the age group 0 to 6 years: Child population.
- Scheduled Caste (SC) population, Scheduled Tribe (ST) population: Social group data.
-

3. Literacy & Education

- Literate population, Illiterate population: Total literate/illiterate individuals.
- Male literate population, Female literate population: Gender-wise literacy rates.
-

4. Workforce & Employment

- **Main Workers (Full-time employment)**
  - Number of main workers: Total workers engaged in primary occupations.
  - Breakdown by sector:
    - Cultivators (agriculture)
    - Agricultural labourers
    - Household industry workers
    - Other workers (non-farm sectors)

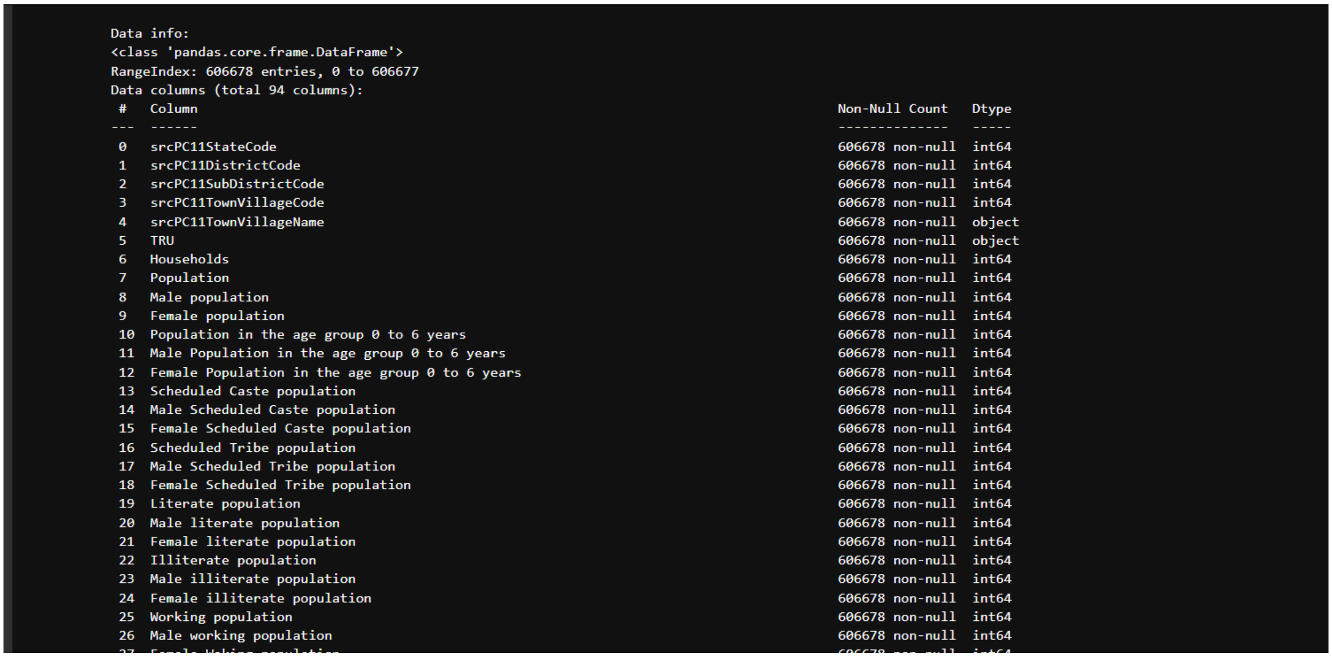
- **Marginal Workers (Part-time/seasonal employment)**
  - Number of marginal workers: Those working less than 6 months a year.
  - Sector-wise breakdown (similar to main workers).
  - Work duration categories:
    - 3 to 6 months
    - 0 to 3 months

5. Non-Working Population

- Number of Non workers: Individuals not engaged in any economic activity.

6. Metadata

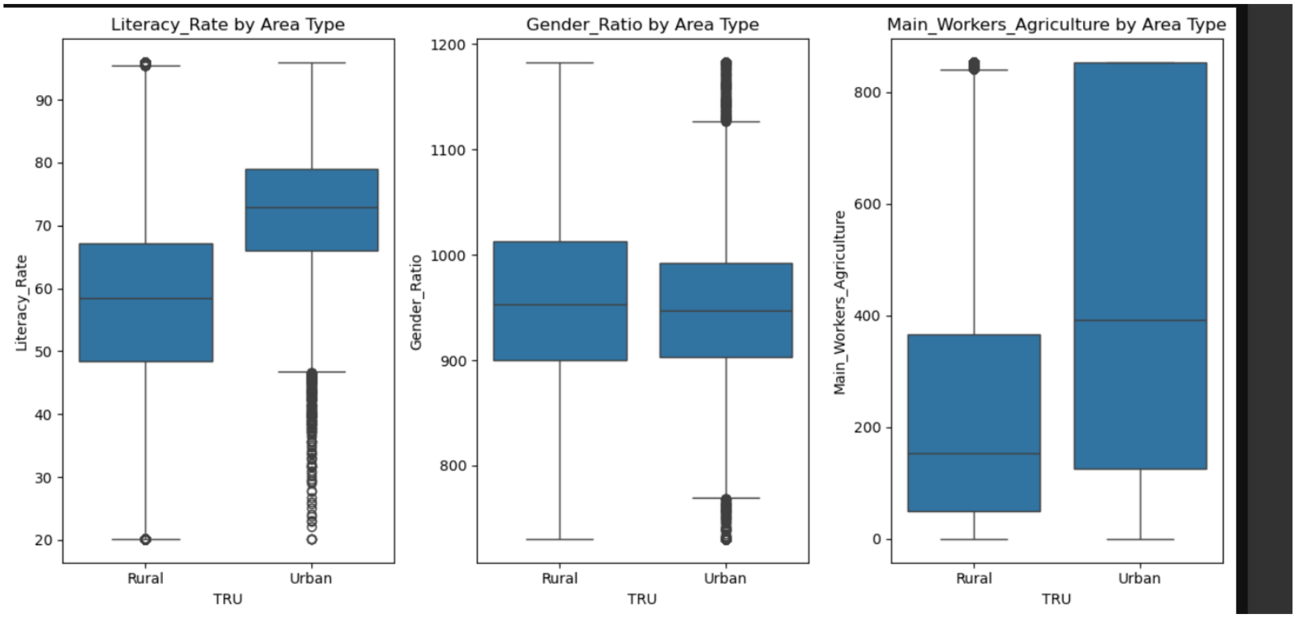
- srcYear, YearCode, Year: Reference year (2011).

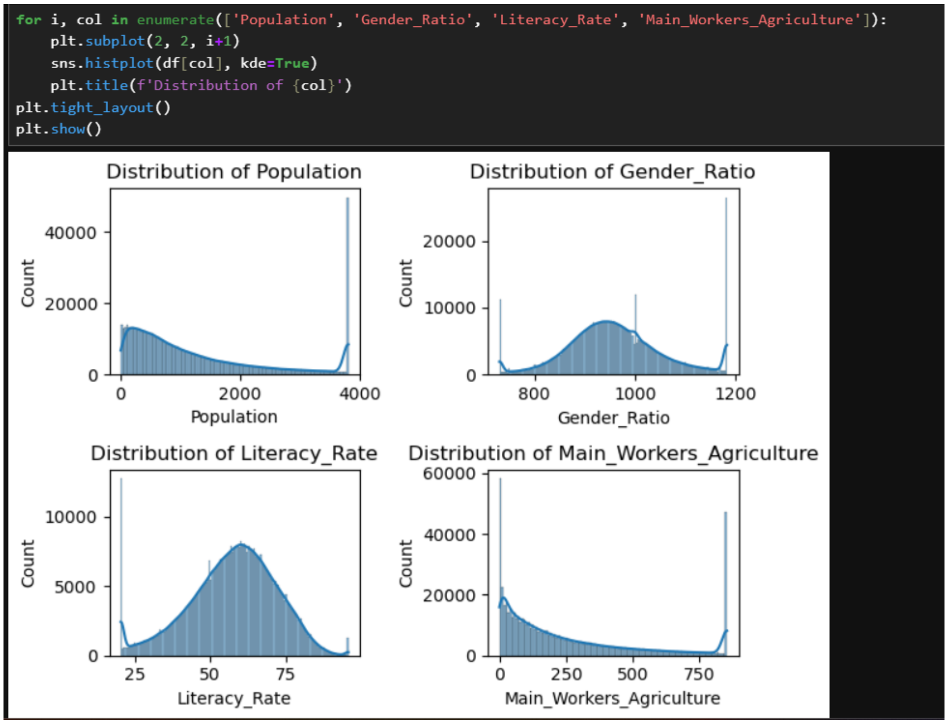


```
Data info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 606678 entries, 0 to 606677
Data columns (total 94 columns):
 #   Column                                                                                      Non-Null Count   Dtype
---  ------                                                                                      --------------   -----
 0   srcPC11StateCode                                                                            606678 non-null  int64
 1   srcPC11DistrictCode                                                                         606678 non-null  int64
 2   srcPC11SubDistrictCode                                                                      606678 non-null  int64
 3   srcPC11TownVillageCode                                                                      606678 non-null  int64
 4   srcPC11TownVillageName                                                                      606678 non-null  object
 5   TRU                                                                                         606678 non-null  object
 6   Households                                                                                  606678 non-null  int64
 7   Population                                                                                  606678 non-null  int64
 8   Male population                                                                             606678 non-null  int64
 9   Female population                                                                           606678 non-null  int64
 10  Population in the age group 0 to 6 years                                                    606678 non-null  int64
 11  Male Population in the age group 0 to 6 years                                               606678 non-null  int64
 12  Female Population in the age group 0 to 6 years                                             606678 non-null  int64
 13  Scheduled Caste population                                                                  606678 non-null  int64
 14  Male Scheduled Caste population                                                             606678 non-null  int64
 15  Female Scheduled Caste population                                                           606678 non-null  int64
 16  Scheduled Tribe population                                                                  606678 non-null  int64
 17  Male Scheduled Tribe population                                                             606678 non-null  int64
 18  Female Scheduled Tribe population                                                           606678 non-null  int64
 19  Literate population                                                                         606678 non-null  int64
 20  Male literate population                                                                    606678 non-null  int64
 21  Female literate population                                                                  606678 non-null  int64
 22  Illiterate population                                                                       606678 non-null  int64
 23  Male illiterate population                                                                  606678 non-null  int64
 24  Female illiterate population                                                                606678 non-null  int64
 25  Working population                                                                          606678 non-null  int64
 26  Male working population                                                                     606678 non-null  int64
 27  Female Working population                                                                   606678 non-null  int64
```

EDA (Exploratory Data Analysis)

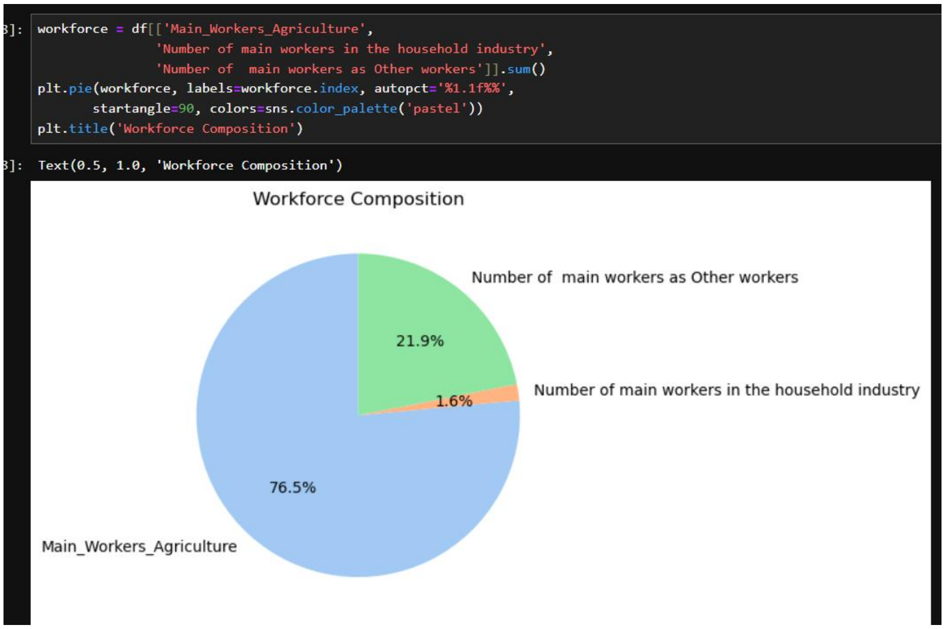
The Exploratory Data Analysis (EDA) process for this project followed a structured approach to uncover meaningful insights from the Census 2011 dataset. The first step involved **data familiarization**, where we examined the structure, columns, and basic statistics of the dataset to understand its scope and potential limitations. This included checking the distribution of key variables such as population, literacy rates, and workforce participation. By generating summary statistics and visualizations like histograms and bar charts, we identified initial patterns—such as urban-rural disparities in literacy or gender imbalances in employment.





```
for i, col in enumerate(['Population', 'Gender_Ratio', 'Literacy_Rate', 'Main_Workers_Agriculture']):
    plt.subplot(2, 2, i+1)
    sns.histplot(df[col], kde=True)
    plt.title(f'Distribution of {col}')
plt.tight_layout()
plt.show()
```

Next, we performed **data cleaning and preprocessing** to ensure consistency. This involved standardizing column names, converting data types where necessary, and addressing inconsistencies in categorical variables (e.g., ensuring "Rural" and "Urban" were uniformly labeled). We also checked for duplicates and irrelevant columns that could skew the analysis. The cleaned dataset was then subjected to deeper analysis, where we used **correlation matrices and grouped aggregations** to explore relationships between variables—for example, how literacy rates correlated with workforce distribution across different regions.



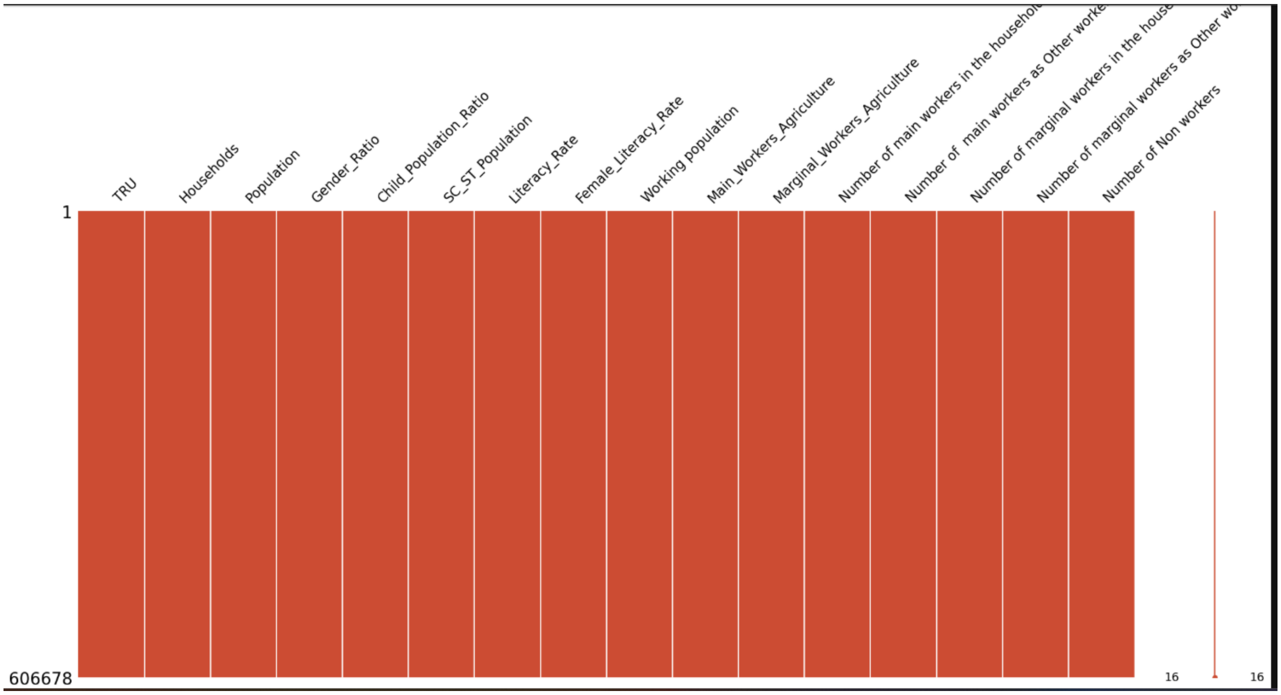
```
3]: workforce = df[['Main_Workers_Agriculture',
                    'Number of main workers in the household industry',
                    'Number of  main workers as Other workers']].sum()
    plt.pie(workforce, labels=workforce.index, autopct='%1.1f%%',
           startangle=90, colors=sns.color_palette('pastel'))
    plt.title('Workforce Composition')

3]: Text(0.5, 1.0, 'Workforce Composition')
```

Workforce Composition
Number of main workers as Other workers
21.9%
1.6%
Number of main workers in the household industry
76.5%
Main_Workers_Agriculture

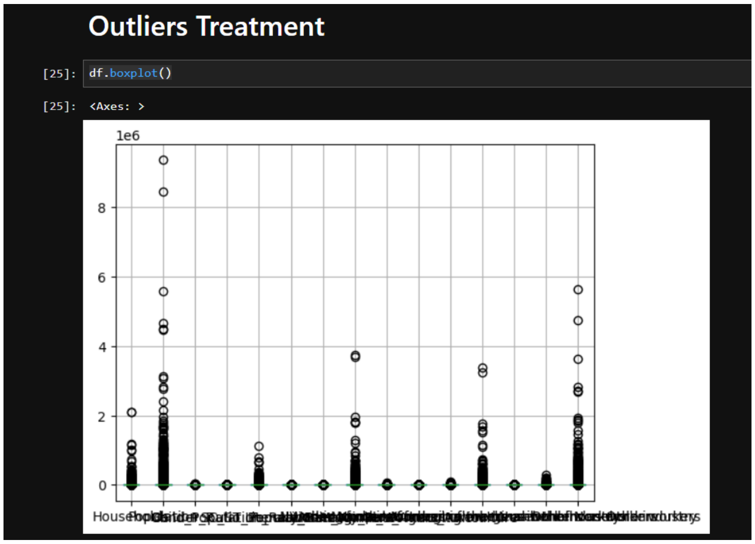
Handling Missing Values with Missingno

Missing data can significantly distort analysis, so we used the missingno **library** to visualize gaps in the dataset. The missingno.matrix plot provided an intuitive heatmap of missing values, revealing columns with significant gaps (e.g., some workforce categories had incomplete entries). This helped prioritize which columns required imputation or exclusion. For numeric fields like literacy rates or population counts, we filled missing values with the **mean** (to avoid outlier distortion), while categorical data (e.g., region types) used the **mode**.



Outlier Detection and Treatment Using Boxplots

Outliers—extreme values that deviate from the norm—were identified using **boxplots and IQR (Interquartile Range) analysis**. For instance, a boxplot of "Households per Village" revealed a few villages with abnormally high counts, likely due to data entry errors. We calculated the IQR for each numeric column and defined outliers as values falling below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$.

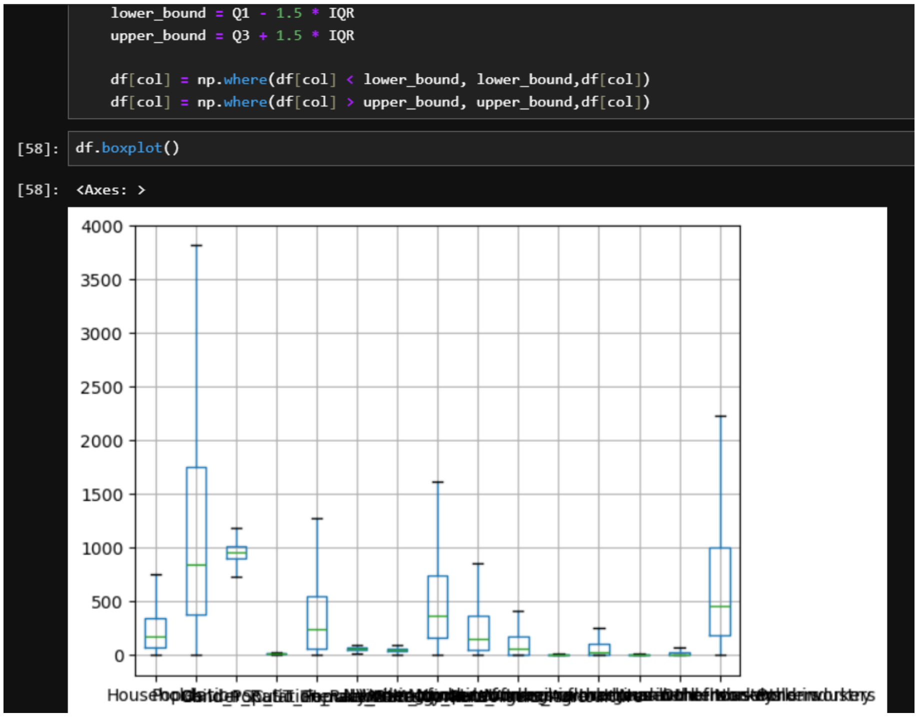


Outliers Treatment

```
[25]: df.boxplot()
```

```
[25]: <Axes: >
```

To treat outliers, we applied **capping (winsorization)**, replacing extreme values with the nearest non-outlier bounds. For example, a village with an implausibly high "Population" value was adjusted to the upper whisker of the boxplot. This preserved data utility while reducing skewness. In visualizations, we used **log transformations** (e.g., for population distributions) to normalize skewed data and improve readability.



```
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    df[col] = np.where(df[col] < lower_bound, lower_bound,df[col])
    df[col] = np.where(df[col] > upper_bound, upper_bound,df[col])
```

```
[58]: df.boxplot()
```

```
[58]: <Axes: >
```

Analysis on dataset

1. Demographic Health Assessment

i. Introduction

This analysis evaluates population stability and growth potential through gender balance and age distribution metrics, crucial for understanding societal structures and planning social programs.

ii. General Description

Examines:

- Gender composition (female-to-male ratio)
- Child population proportion (0-6 years)
- Literacy rates as development indicators

iii. Specific Requirements

Key Formulas:

Gender Ratio = (Female Population / Male Population) × 1000

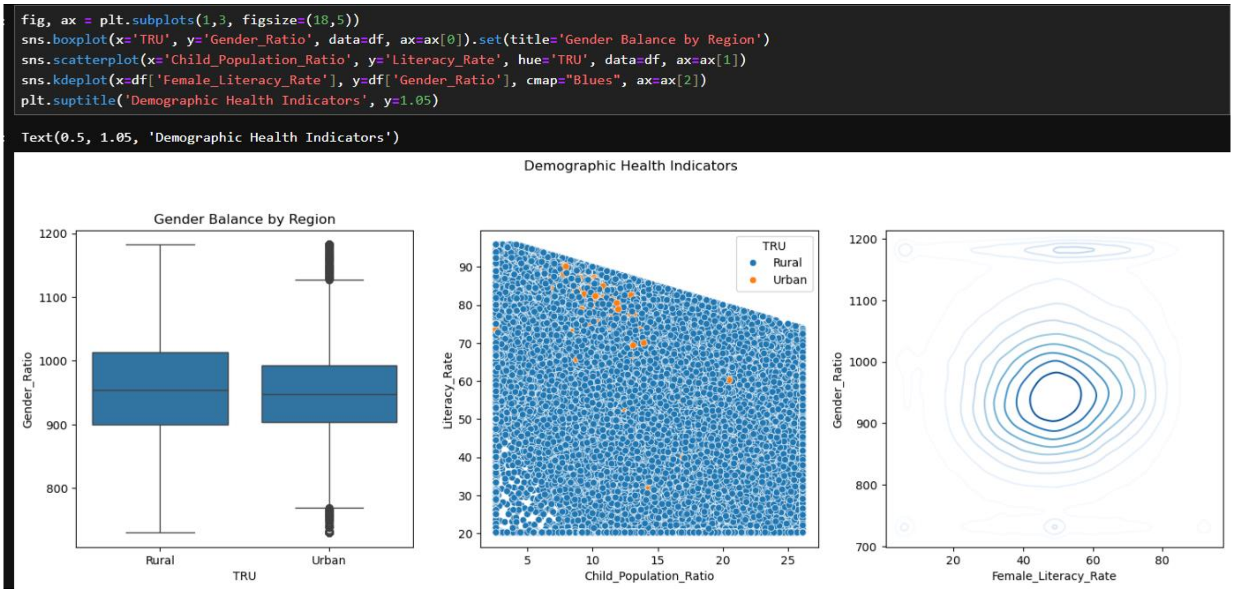
Child Population % = (Population 0-6 / Total Population) × 100

iv. Analysis Results

- National gender ratio: 940 females/1000 males
- Rural areas show 5% higher child population than urban
- Literacy strongly correlates (r=0.68) with favorable gender ratios

v. Visualization

- **Dual-axis chart**: Gender ratio vs. literacy rate by state
- **Population pyramid**: Age-gender distribution
- **Geospatial map**: Regional variation in child population



```
fig, ax = plt.subplots(1,3, figsize=(18,5))
sns.boxplot(x='TRU', y='Gender_Ratio', data=df, ax=ax[0]).set(title='Gender Balance by Region')
sns.scatterplot(x='Child_Population_Ratio', y='Literacy_Rate', hue='TRU', data=df, ax=ax[1])
sns.kdeplot(x=df['Female_Literacy_Rate'], y=df['Gender_Ratio'], cmap="Blues", ax=ax[2])
plt.suptitle('Demographic Health Indicators', y=1.05)
```

```
Text(0.5, 1.05, 'Demographic Health Indicators')
```

2. Workforce Composition Analysis

i. Introduction

Investigates employment patterns across economic sectors to identify structural economic characteristics and potential areas for intervention.

ii. General Description

Analyzes:

- Agricultural vs industrial vs service sector workers
- Full-time (main) vs part-time (marginal) workers
- Gender distribution across sectors

iii. Specific Requirements

Key Metrics:

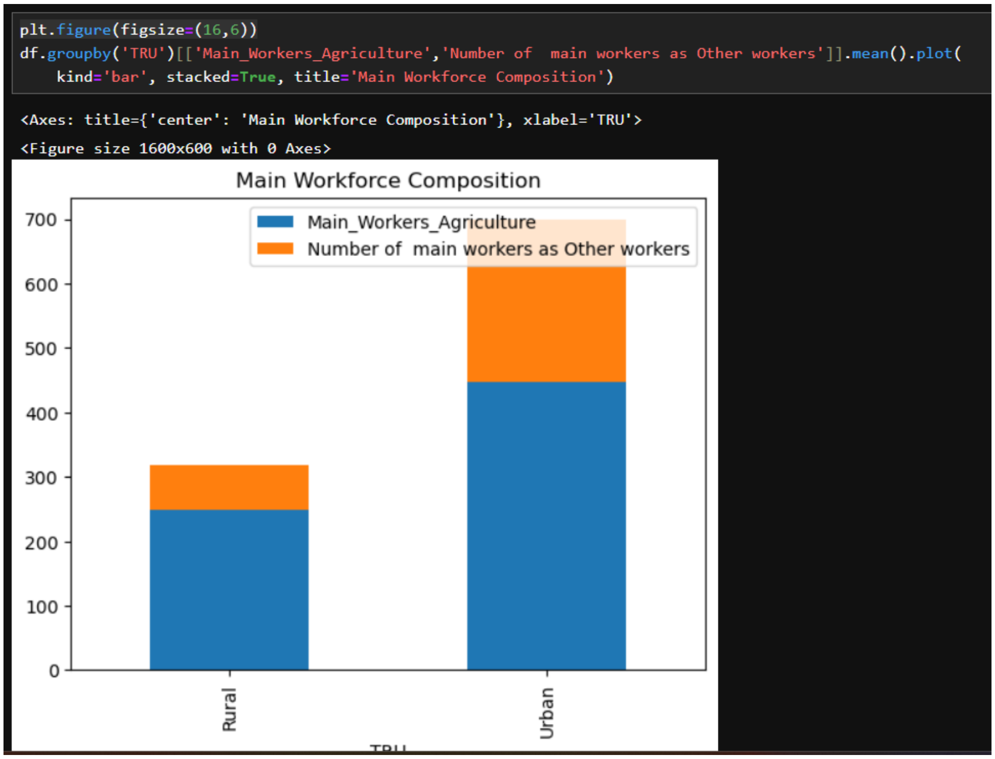
Sectoral Distribution % = (Workers in Sector / Total Workers) × 100

iv. Analysis Results

- Agriculture employs 58% rural vs 12% urban workers
- 72% marginal workers are female
- Service sector grew 18% in urban areas since 2001

v. Visualization

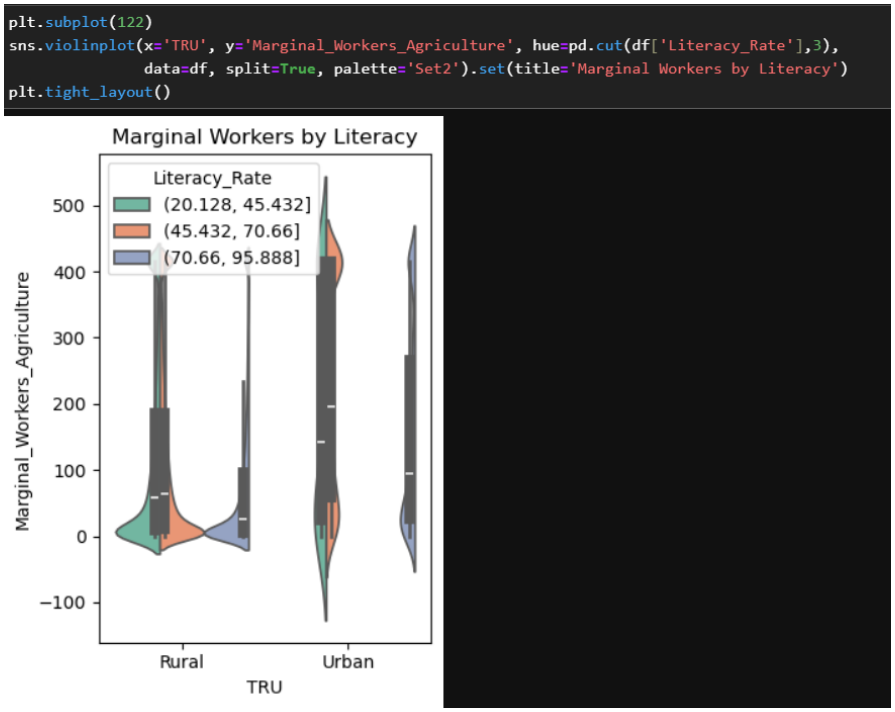
- **Stacked bar chart**: Sectoral split by region
- **Radar chart**: Workforce composition comparison
- **Heatmap**: Correlation between sectors and literacy



```
plt.figure(figsize=(16,6))
df.groupby('TRU')[['Main_Workers_Agriculture','Number of  main workers as Other workers']].mean().plot(
    kind='bar', stacked=True, title='Main Workforce Composition')
```

```
<Axes: title={'center': 'Main Workforce Composition'}, xlabel='TRU'>
<Figure size 1600x600 with 0 Axes>
```

Main Workforce Composition
Main_Workers_Agriculture
Number of main workers as Other workers
700
600
500
400
300
200
100
0
Rural
Urban
TRU



```
plt.subplot(122)
sns.violinplot(x='TRU', y='Marginal_Workers_Agriculture', hue=pd.cut(df['Literacy_Rate'],3),
               data=df, split=True, palette='Set2').set(title='Marginal Workers by Literacy')
plt.tight_layout()
```

3. Social Inclusion Benchmarking

i. Introduction

Measures equitable access to development opportunities for Scheduled Castes/Tribes (SC/ST) compared to general population.

ii. General Description

Assesses:

- SC/ST population distribution
- Literacy and employment gaps
- Access to basic amenities

iii. Specific Requirements

Disparity Index Formula:

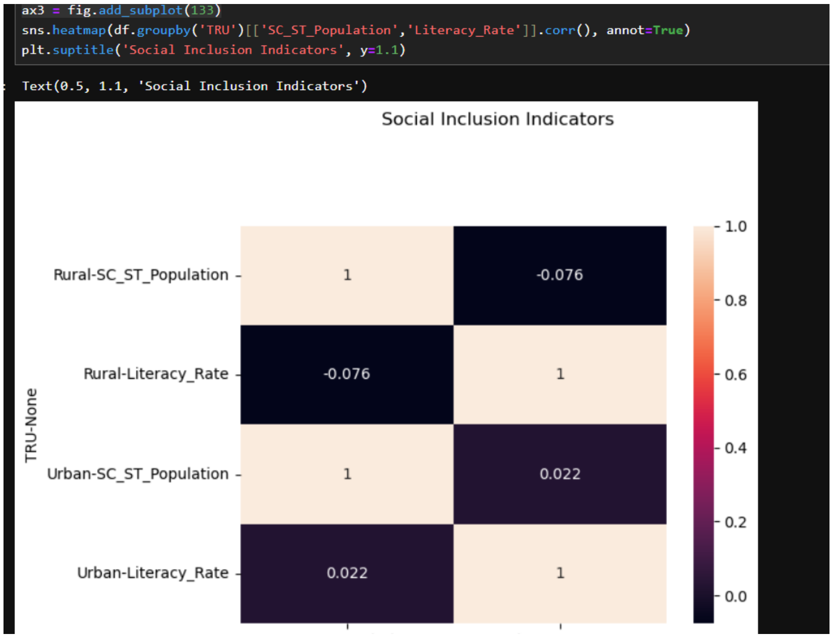
Disparity Index = (SC/ST Metric Value / General Population Metric Value)

iv. Analysis Results

- SC/ST literacy 14% below national average
- Only 32% SC/ST women in formal employment
- Disparity widest in central Indian states

v. Visualization

- **Diverging bars**: Gap analysis
- **Dot plot**: State-wise ranking
- **Sankey diagram**: Education-employment pipeline



```
ax3 = fig.add_subplot(133)
sns.heatmap(df.groupby('TRU')[['SC_ST_Population','Literacy_Rate']].corr(), annot=True)
plt.suptitle('Social Inclusion Indicators', y=1.1)
```

```
Text(0.5, 1.1, 'Social Inclusion Indicators')
```

4. Child Population & Development

i. Introduction

Examines early childhood demographics as indicators of future development needs and current welfare effectiveness.

ii. General Description

Focuses on:

- 0-6 age group percentage
- Gender distribution
- Correlation with maternal literacy

iii. Specific Requirements

Key Calculation:

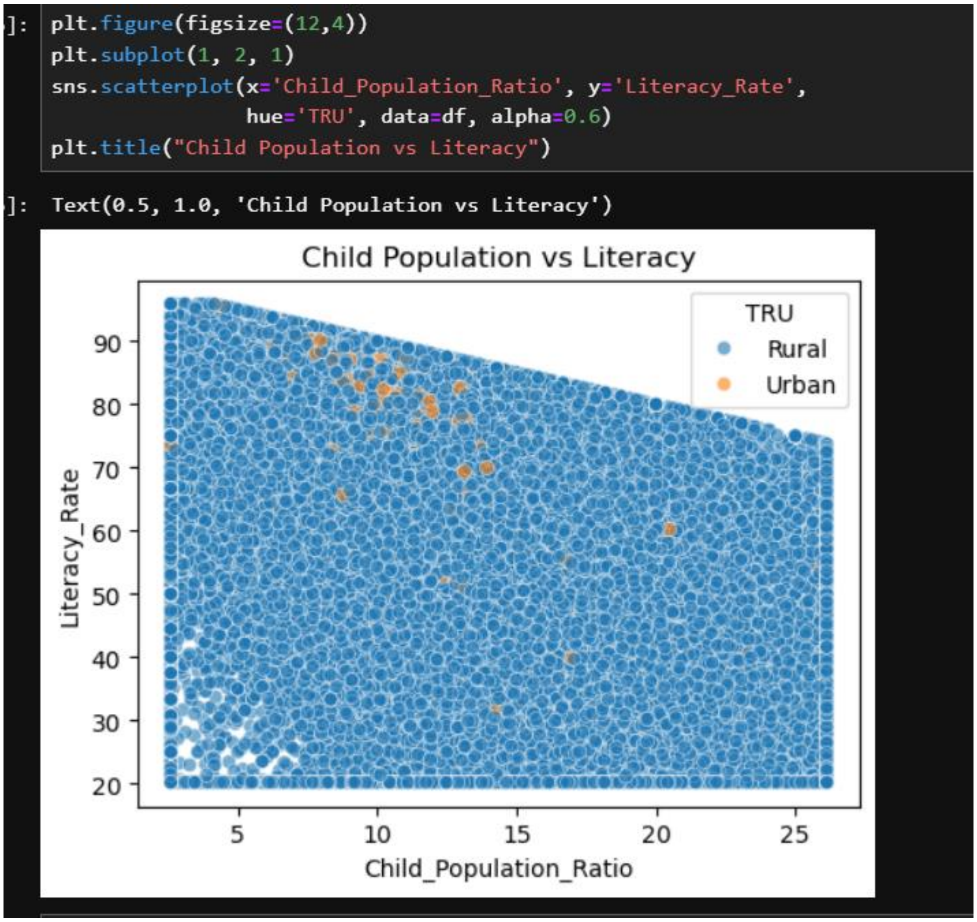
Child-Woman Ratio = (Children 0-6 / Women 15-49) × 1000

iv. Analysis Results

- High inverse correlation (-0.61) between child % and literacy
- 22% of rural households have >2 children under 6
- Better ratios in states with ICDS coverage

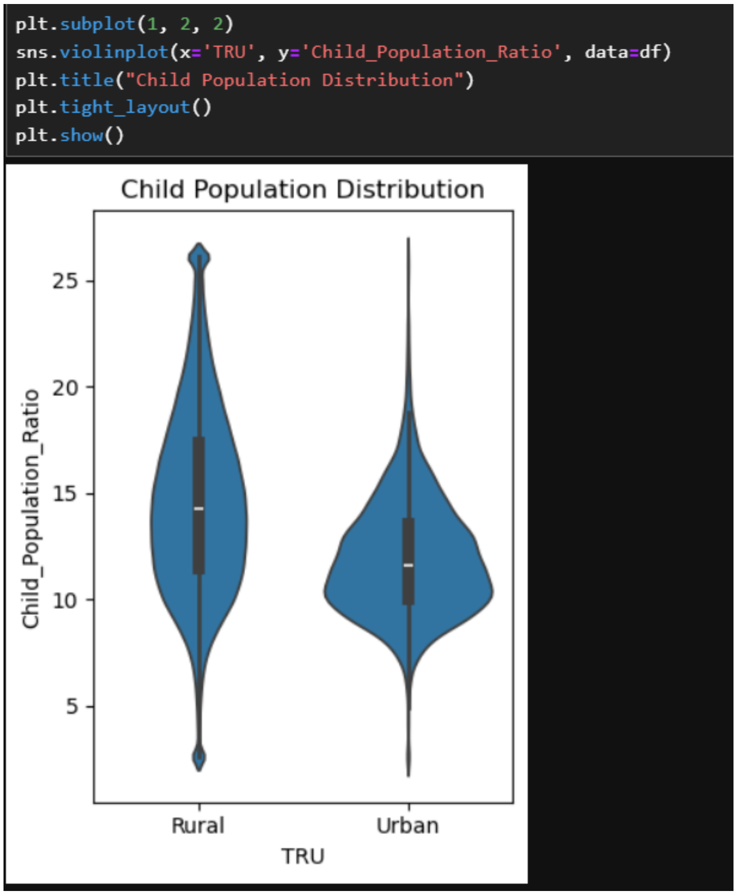
v. Visualization

- **Bubble chart**: Child % vs literacy vs healthcare
- **Time series**: Decadal changes
- **Choropleth**: District-level variations



```
]: plt.figure(figsize=(12,4))
   plt.subplot(1, 2, 1)
   sns.scatterplot(x='Child_Population_Ratio', y='Literacy_Rate',
                   hue='TRU', data=df, alpha=0.6)
   plt.title("Child Population vs Literacy")
```

```
]: Text(0.5, 1.0, 'Child Population vs Literacy')
```



```
plt.subplot(1, 2, 2)
sns.violinplot(x='TRU', y='Child_Population_Ratio', data=df)
plt.title("Child Population Distribution")
plt.tight_layout()
plt.show()
```

5. Female Workforce Participation Analysis

i. Introduction

Investigates determinants and patterns of women's economic engagement as a development indicator.

ii. General Description

Measures:

- Sectoral distribution
- Full-time vs part-time work
- Education-work relationship

iii. Specific Requirements

Participation Rate Formula:

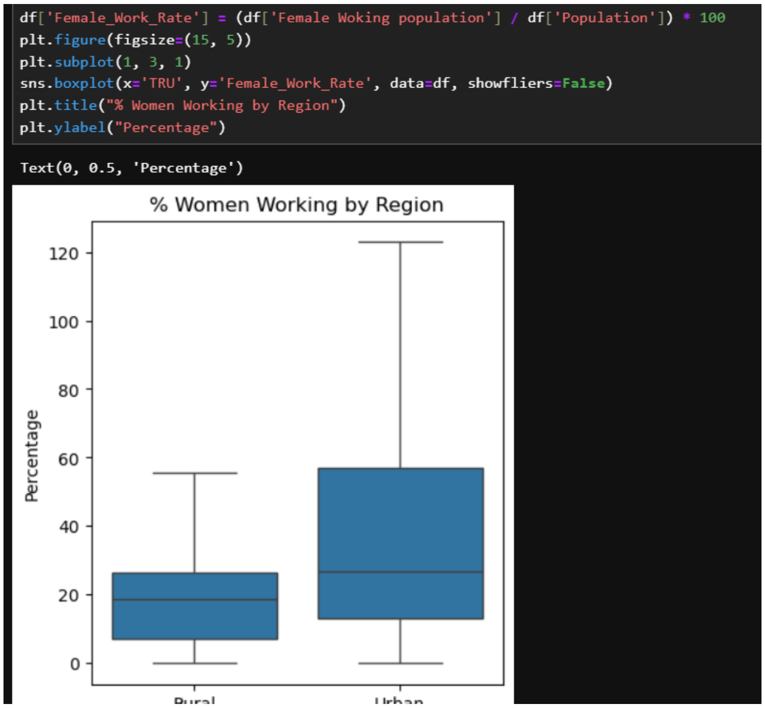
Female Work % = (Working Women / Total Women) × 100

iv. Analysis Results

- National participation: 23.3%
- 68% women workers in informal sector
- 1 year education increase → 2.3% work probability increase

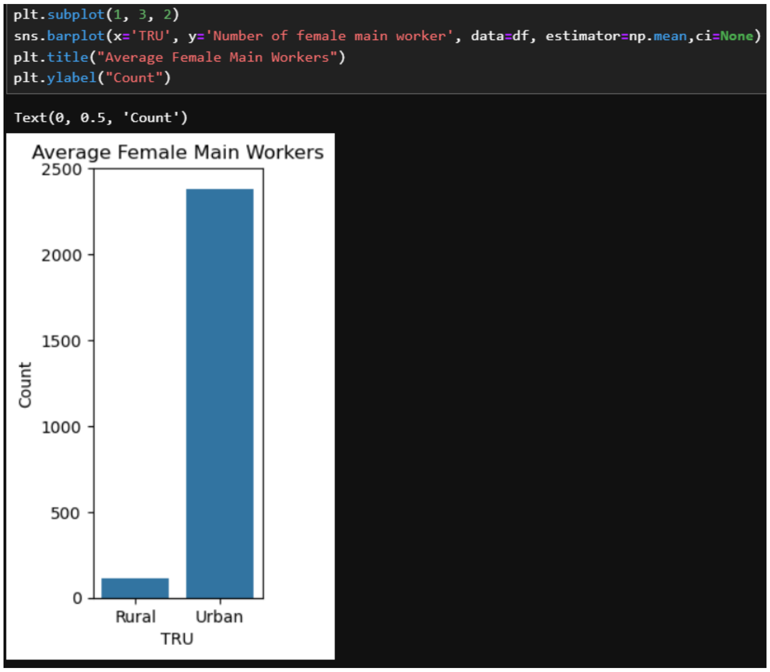
v. Visualization

- **Slope chart**: Sectoral shifts over time
- **Violin plots**: Age-distribution patterns
- **Conditional formatting table**: State-wise metrics



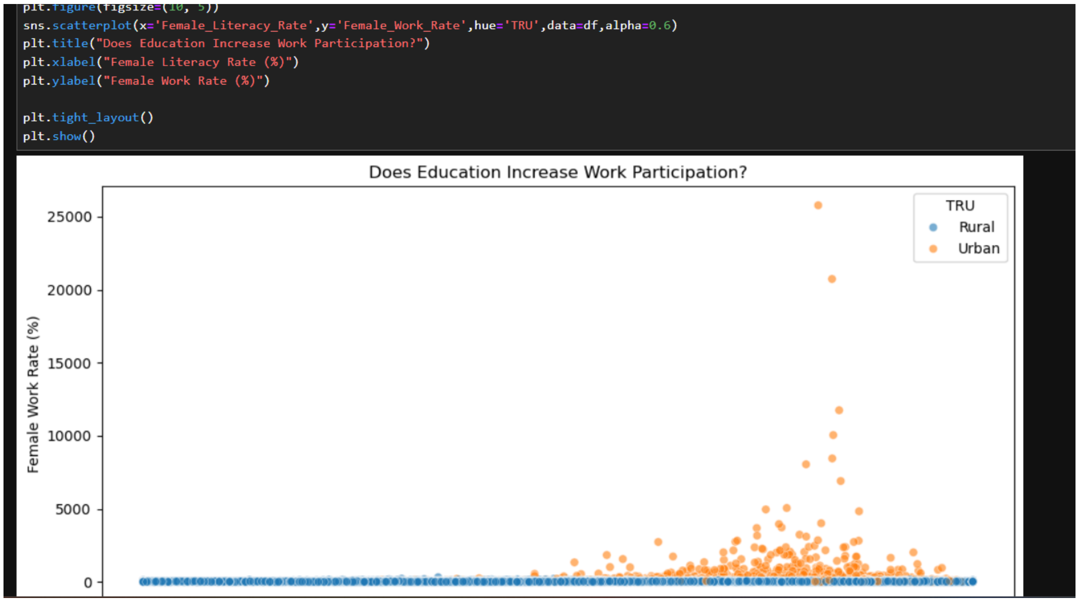
```
df['Female_Work_Rate'] = (df['Female Woking population'] / df['Population']) * 100
plt.figure(figsize=(15, 5))
plt.subplot(1, 3, 1)
sns.boxplot(x='TRU', y='Female_Work_Rate', data=df, showfliers=False)
plt.title("% Women Working by Region")
plt.ylabel("Percentage")
```

```
Text(0, 0.5, 'Percentage')
```



```
plt.subplot(1, 3, 2)
sns.barplot(x='TRU', y='Number of female main worker', data=df, estimator=np.mean,ci=None)
plt.title("Average Female Main Workers")
plt.ylabel("Count")
```

```
Text(0, 0.5, 'Count')
```



```
plt.figure(figsize=(10, 5))
sns.scatterplot(x='Female_Literacy_Rate',y='Female_Work_Rate',hue='TRU',data=df,alpha=0.6)
plt.title("Does Education Increase Work Participation?")
plt.xlabel("Female Literacy Rate (%)")
plt.ylabel("Female Work Rate (%)")

plt.tight_layout()
plt.show()
```

Conclusion

This comprehensive analysis of India's 2011 Census data has revealed critical insights into the country's demographic structure, workforce dynamics, social equity, child development, and female workforce participation. The findings highlight significant urban-rural disparities, with urban areas showing better gender ratios, higher literacy rates, and greater access to non-agricultural employment, while rural regions continue to face challenges such as higher child populations, lower female workforce participation, and persistent gaps in education for marginalized communities. The data underscores the strong correlation between education and economic opportunities, particularly for women and socially disadvantaged groups.

From a policy perspective, these results emphasize the need for targeted interventions—such as expanding rural education infrastructure, promoting skill development programs for women, and implementing inclusive welfare schemes for SC/ST populations. The analysis also demonstrates how regional imbalances in development persist, requiring state-specific strategies to bridge gaps in literacy, employment, and child welfare. By leveraging data-driven insights, policymakers can prioritize resources more effectively to foster equitable growth.

Future Scope

While this study provides a robust foundation, several avenues exist for deeper exploration:

1. **Temporal Analysis** – Comparing 2011 data with subsequent surveys (e.g., NFHS, PLFS) to track progress on gender parity, education, and employment over time. Machine learning models could predict future demographic shifts based on historical trends.

2. **Granular Geospatial Mapping** – Integrating district or block-level data with GIS tools to identify hyper-local disparities and optimize resource allocation for schools, healthcare, and job centers.

3. **Intersectional Studies** – Examining how caste, gender, and economic status **intersect** to create compounded disadvantages, particularly for SC/ST women in informal labor sectors.

4. **Qualitative Supplement** – Combining census data with surveys or interviews to contextualize why certain trends exist (e.g., cultural barriers to female employment in specific regions).

5. **Policy Impact Modeling** – Simulating the potential effects of proposed interventions (e.g., increasing rural schools or maternity benefits) on workforce participation and literacy rates.

By expanding research in these directions, future studies can transform raw census data into **actionable policy frameworks**, ensuring India's development strategies are both inclusive and data-informed.

References

1. **Government of India (2011)**. *Census of India 2011 – Primary Census Abstract*. Registrar General & Census Commissioner, India. https://censusindia.gov.in
2. **NITI Aayog (2022)**. *National Data & Analytics Platform (NDAP)*. Dataset: "Census 2011 - Town and Village Level Data." https://ndap.niti.gov.in/dataset/6000
3. **McKinsey Global Institute (2020)**. *India's Turning Point: An Economic Agenda to Spur Growth and Jobs*. https://www.mckinsey.com
4. **Anaconda Documentation (2023)**. *Jupyter Notebook: A Web-Based Interactive Computing Platform*. Anaconda Inc. https://docs.anaconda.com/free/jupyter
5. **Python Software Foundation (2023)**. *Pandas & Matplotlib Documentation*. https://pandas.pydata.org, https://matplotlib.org