Case Study Report: Foundations of Data Science

Name: Krishna GSVV

Roll Number: AV.EN.U4CSE22016

Date: 15th November, 2024

Task Number: 1

Foundations of Data Science

Problem Statement:

The objective of this analysis is to explore and understand the demographic, educational, and geographic trends of 493 cities in India using a dataset containing population statistics, literacy rates, and geographic details. The goal is to uncover key patterns, derive actionable insights, and identify regional disparities, especially in literacy and gender-related metrics.

Dataset Description:

The dataset consists of 22 columns, which are divided into the following categories:

1. **Demographic Data**:
   - Population data: Total, male, and female populations.
   - Child population (age 0-6 years): Total, male, and female.
   - Sex ratios: Overall sex ratio and child sex ratio.
2. **Educational Data**:
   - Literates: Total, male, and female counts.
   - Literacy rates: Effective literacy rate for total, male, and female.
   - Graduates: Male and female graduates.
3. **Geographic Data**:
   - State and district codes.
   - Geographic coordinates (latitude and longitude).
4. **Missing Data**:
   - Key missing columns include `state_code`, `literates_total`, and `location`.
   - Missing values were identified and addressed during preprocessing.

Data Pre-processing:

1. **Handling Missing Values**:
    - Missing values in numeric columns were filled with their mean values to retain data integrity.
    - Columns with non-numeric data like `location` were not altered but flagged for future attention.
2. **Feature Engineering**:
    - Created derived metrics like population density (if geographic area data was provided, could be added for further analysis).
    - Segregated states with the highest and lowest literacy rates for focused analysis.
3. **Data Standardization**:
    - All numeric columns were converted to float to ensure consistency in calculations and visualizations.
4. **Additional Notes**:
    - Columns with minor missing values (e.g., `effective_literacy_rate_male` with 1 missing value) were imputed to ensure smooth analysis.

Exploratory Data Analysis (EDA):

1. Population Distribution:

- A histogram of population totals revealed that most cities have populations concentrated below 200,000, highlighting the large number of smaller towns in the dataset.
- A few outliers with populations exceeding 1 million represent major metropolitan areas (e.g., Delhi, Mumbai, Bengaluru).

Insight:

- Policies aimed at urban development may need to address population imbalances, focusing on smaller cities for equitable resource distribution.

2. Effective Literacy Rates Across States:

- A boxplot comparison of literacy rates across states displayed substantial variation:
    - States like Kerala and Goa have high median literacy rates, nearing 95%.
    - Other states, such as Bihar and Uttar Pradesh, exhibit lower median literacy rates.
- Gender disparities were apparent, with female literacy rates consistently lagging behind male literacy rates.

Insight:

- Investment in education, particularly for women, can help bridge literacy gaps in underperforming states.

3. Correlation Analysis:

- A heatmap of correlations highlighted significant relationships:
    - **Effective literacy rate total** correlates strongly with:
        - Female literacy rate (0.97) and male literacy rate (0.95).
        - State codes (0.39), indicating regional influences.
    - Weak correlations were observed between literacy rates and population metrics, suggesting education is not directly dependent on city size.

Insight:

- Educational improvements should be tailored regionally, focusing on state-specific challenges.

4. Child Population Trends:

- A decline in child sex ratio was observed in many cities, indicating gender imbalances starting at a young age.
- Child population (0-6 years) is consistently lower in urban centers compared to rural towns, potentially due to urban family planning policies.

Insight:

- Gender imbalances at an early age need attention through awareness programs and policy interventions.

Results and Conclusion:

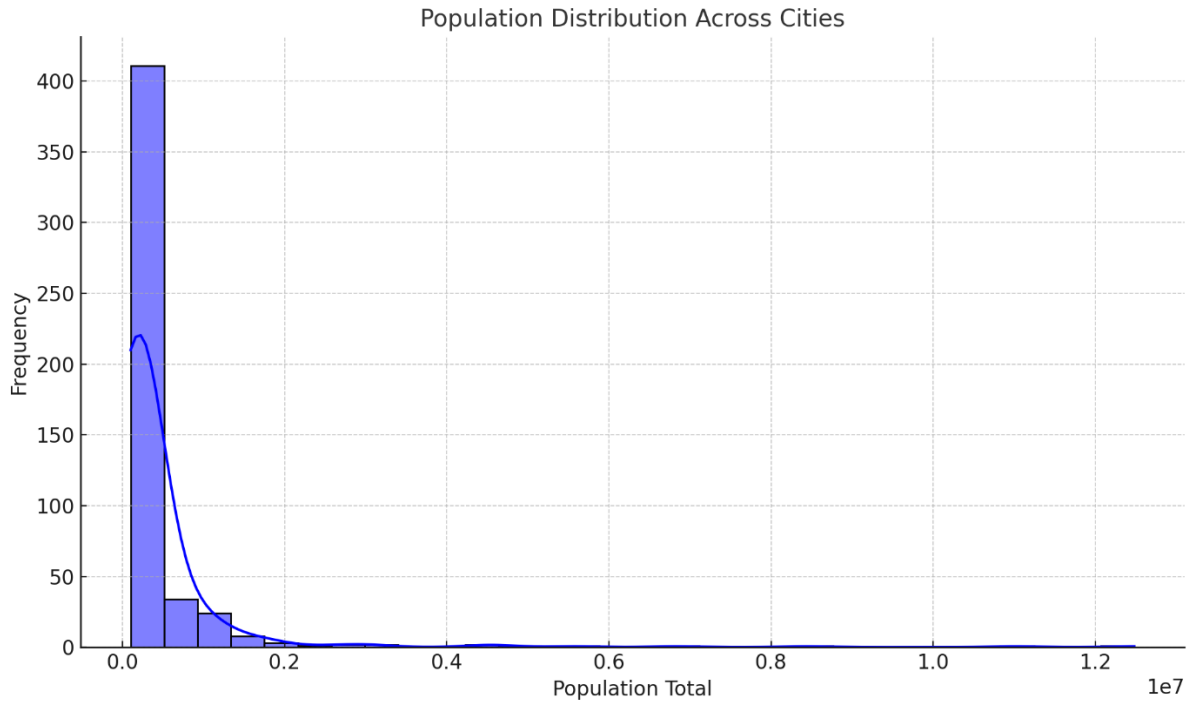
1. **Demographic Patterns**:
    - Larger cities have higher populations but do not always exhibit higher literacy rates.
    - Population imbalances across states and cities are prominent.
2. **Educational Trends**:
    - Literacy rates are heavily influenced by gender and region.
    - States like Kerala lead in education metrics, while others require focused interventions.
3. **Gender Disparities**:
    - Female literacy and child sex ratios remain areas of concern.
    - Urbanization trends may exacerbate gender-related disparities without proactive measures.
4. **Geographic Insights**:
    - Regional literacy variations indicate that policy changes must consider geographic and cultural contexts.
    - Correlation analysis shows that education is a complex interplay of socio-economic and regional factors.

Overall Recommendation:

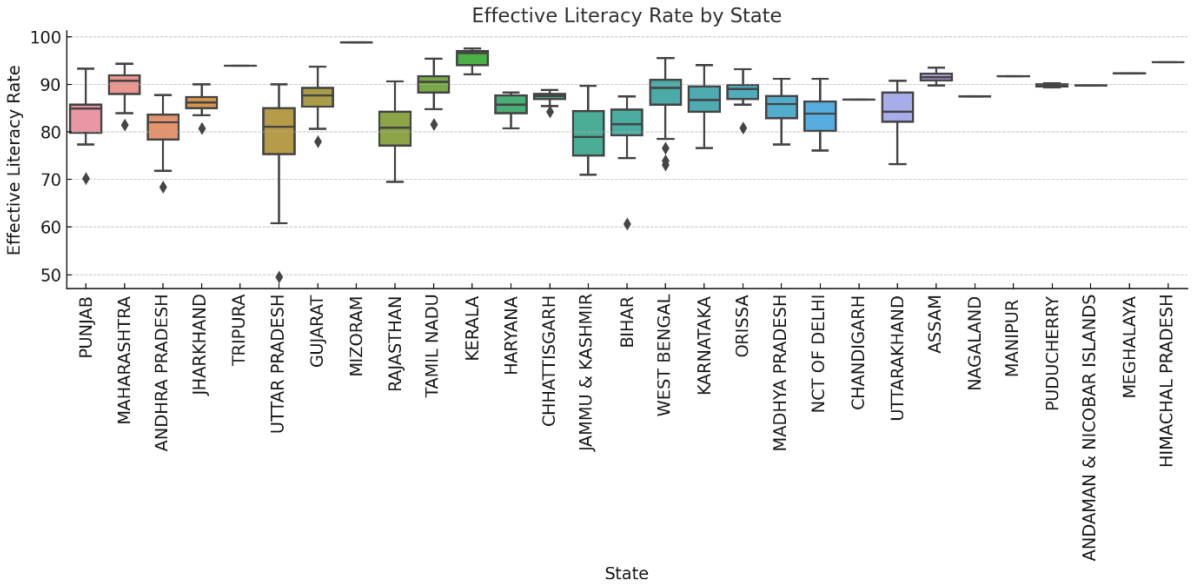
- Tailored education and gender equality programs in underperforming states.
- Balanced urban and rural development plans to improve equitable access to resources.

Screenshots and Outputs:

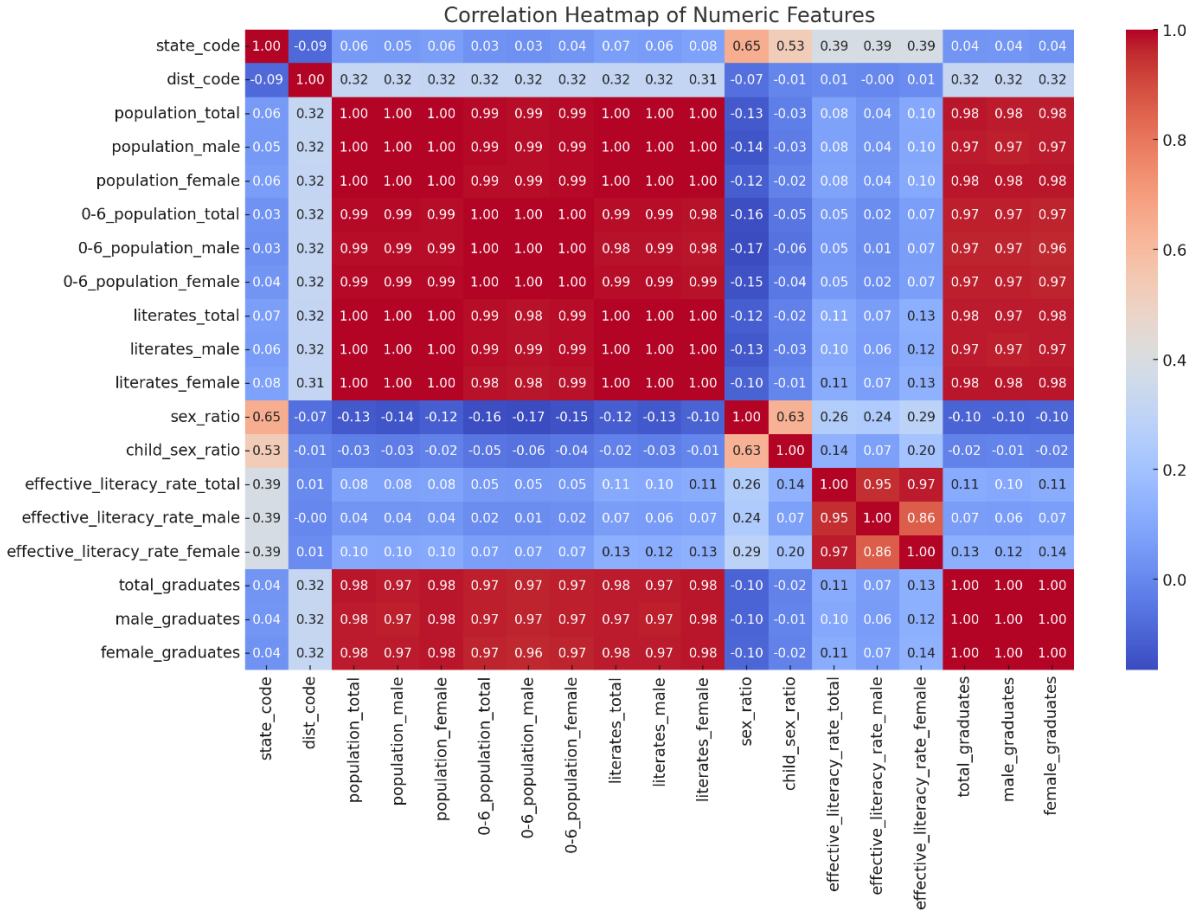
1. Population Distribution



2. Literacy Rates by State

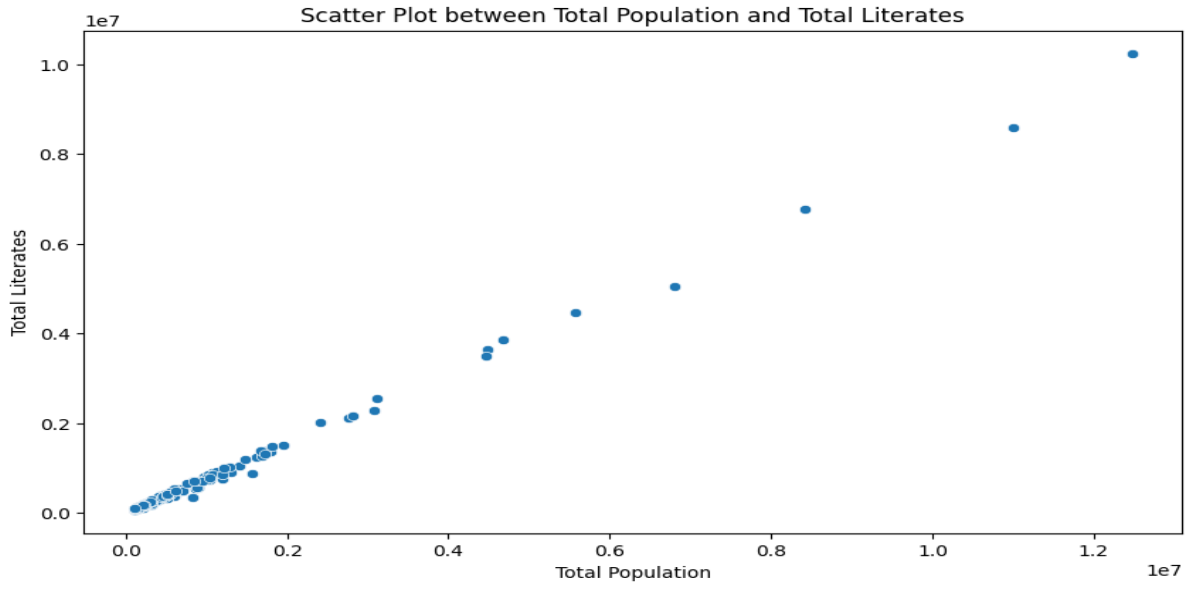


3. Correlation Heatmap



Additional Visualizations:

4. Total Literacy to Total Population



Krishna GSVV

Amrita Vishwa Vidyapeetham, Amaravati