# INTODUCTION TO DATA MANAGEMENT

## PROJECT REPORT

(Project Semester January-April 2025)

# City-wise Population by Mother Tongue- Tamil Nadu


Submitted by

**YUVANESH MANI**

Registration No-12307061

Programme and Section- BTECH(CSE) & K23GF

Course Code -INT217

Under the Guidance of

**Sandeep Kaur**

**Discipline of CSE/IT**


## CERTIFICATE

This is to certify that **Yuvanesh Mani** bearing Registration no. **12307061** has completed INT217 project titled, **"City-wise Population by Mother Tongue- Tamil Nadu"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**
**Designation of the Supervisor**
**School of Engineering,** Lovely
Professional University
Phagwara, Punjab.

Date: 15/04/2024

## DECLARATION

I, **Yuvanesh Mani**, student of BTECH (CSE) 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: **15/04/2025**

Registration No. **12307061**

## ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my project guide, Sandeep Kaur, for his valuable guidance, continuous support, and encouragement throughout the course of this project. His mentorship has helped me grow academically and personally.

I also extend my thanks to the faculty members and peers at Lovely Professional University for providing an environment that encourages learning and innovation.

## TABLE OF CONTENTS

# 1. Introduction

Language is a foundational aspect of culture and communication. Understanding the distribution of **mother tongues** across a region can help inform urban planning, educational resource allocation, and public service delivery.

This project focuses on the **Tamil Nadu region of India**, analysing data from the **2001 Census** related to **city-wise population categorized by mother tongue**. The selected towns for detailed analysis include **Gummidipoondi**, **Arani**, and **Ponneri**.

The objective of the project is to perform **Exploratory Data Analysis (EDA)** using **Microsoft Excel**, leveraging its powerful capabilities like **PivotTables**, **slicers**, **charts**, and **formulas** to uncover patterns in language distribution by gender, area (urban/rural), and town.

# 2. Source of Dataset

## Dataset Origin

- **Source:** Open Government Data (OGD) Platform India – data.gov.in
- **Dataset Title:** City-wise Population by Mother Tongue (2001), Tamil Nadu
- **Format:** CSV file (converted to Excel for analysis)

## Geographical Focus

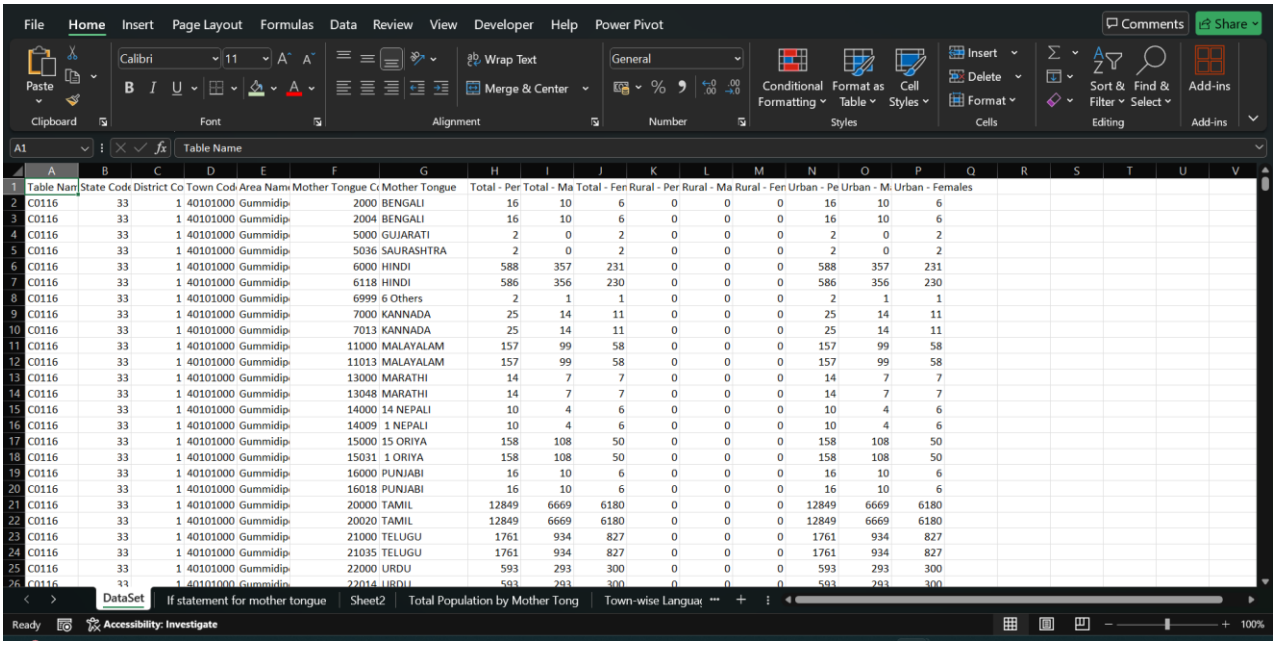
- **State:** Tamil Nadu
- **Towns Analysed:**
  - **Gummidipoondi**
  - **Arani**
  - **Ponneri**

## Description of Data Fields

- **Codes:**
  - State Code, District Code, Town Code
  - Language Code (for each mother tongue)
- **Names:**
  - Town Name (Area Name)
  - Mother Tongue Name
- **Population Split:**
  - Total Persons
  - Total Males
  - Total Females

- Urban Population
- Rural Population

This dataset provides a multidimensional view of population segments, allowing us to analyse not just raw counts but relationships between language, gender, and geographic distribution.



| Table Nam | State Code | District Co | Town Code | Area Name | Mother Tongue Co | Mother Tongue | Total - Per | Total - Ma | Total - Fen | Rural - Per | Rural - Ma | Rural - Fen | Urban - Pe | Urban - M | Urban - Females |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C0116 | 33 | 1 | 40101000 | Gummidip | 2000 | BENGALI | 16 | 10 | 6 | 0 | 0 | 0 | 16 | 10 | 6 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 2004 | BENGALI | 16 | 10 | 6 | 0 | 0 | 0 | 16 | 10 | 6 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 5000 | GUJARATI | 2 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 2 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 5036 | SAURASHTRA | 2 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 2 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 6000 | HINDI | 588 | 357 | 231 | 0 | 0 | 0 | 588 | 357 | 231 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 6118 | HINDI | 586 | 356 | 230 | 0 | 0 | 0 | 586 | 356 | 230 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 6999 | 6 Others | 2 | 1 | 1 | 0 | 0 | 0 | 2 | 1 | 1 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 7000 | KANNADA | 25 | 14 | 11 | 0 | 0 | 0 | 25 | 14 | 11 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 7013 | KANNADA | 25 | 14 | 11 | 0 | 0 | 0 | 25 | 14 | 11 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 11000 | MALAYALAM | 157 | 99 | 58 | 0 | 0 | 0 | 157 | 99 | 58 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 11013 | MALAYALAM | 157 | 99 | 58 | 0 | 0 | 0 | 157 | 99 | 58 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 13000 | MARATHI | 14 | 7 | 7 | 0 | 0 | 0 | 14 | 7 | 7 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 13048 | MARATHI | 14 | 7 | 7 | 0 | 0 | 0 | 14 | 7 | 7 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 14000 | 14 NEPALI | 10 | 4 | 6 | 0 | 0 | 0 | 10 | 4 | 6 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 14009 | 1 NEPALI | 10 | 4 | 6 | 0 | 0 | 0 | 10 | 4 | 6 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 15000 | 15 ORIYA | 158 | 108 | 50 | 0 | 0 | 0 | 158 | 108 | 50 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 15031 | 1 ORIYA | 158 | 108 | 50 | 0 | 0 | 0 | 158 | 108 | 50 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 16000 | PUNJABI | 16 | 10 | 6 | 0 | 0 | 0 | 16 | 10 | 6 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 16018 | PUNJABI | 16 | 10 | 6 | 0 | 0 | 0 | 16 | 10 | 6 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 20000 | TAMIL | 12849 | 6669 | 6180 | 0 | 0 | 0 | 12849 | 6669 | 6180 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 20020 | TAMIL | 12849 | 6669 | 6180 | 0 | 0 | 0 | 12849 | 6669 | 6180 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 21000 | TELUGU | 1761 | 934 | 827 | 0 | 0 | 0 | 1761 | 934 | 827 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 21035 | TELUGU | 1761 | 934 | 827 | 0 | 0 | 0 | 1761 | 934 | 827 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 22000 | URDU | 593 | 293 | 300 | 0 | 0 | 0 | 593 | 293 | 300 |
| C0116 | 33 | 1 | 40101000 | Gummidip | 22014 | URDU | 593 | 293 | 300 | 0 | 0 | 0 | 593 | 293 | 300 |

# 3. EDA Process

## 3.1 Data Import

- Imported CSV into Excel using **Power Query** for clean structure and future refresh capability.
- Verified column mapping and data types during import.

## 3.2 Data Cleaning

- Removed unnecessary whitespace using `TRIM()`
- Renamed headers for clarity (e.g., "Mother Tongue Name" → "Language")
- Handled duplicate entries (e.g., same language with multiple codes) using `IF()` and `SUBSTITUTE()`
- Removed rows with missing or null population values

## 3.3 Data Transformation

- Created **calculated columns**:
  - `% Share in Town` = (Language Population / Total Town Population) × 100
  - `Gender Ratio` = Males / Females

- Created helper columns for **area classification**, language grouping
- Merged records with the same town and language for consistent aggregation

### 3.4 Sorting and Filtering

- Applied custom sort orders:
  - By total population (descending)
  - Alphabetical by language
  - Grouped by urban/rural

### 3.5 Data Validation

- Used `ISNUMBER()` to check numeric integrity
- Applied **conditional formatting** to highlight invalid entries
- Filtered for preview summaries before deep analysis

# 4. Analysis on Dataset

### Analysis 1: Most Spoken Languages by Total Population

**i. Introduction:**
Identify which languages are spoken by the largest populations across selected towns.

**ii. Description:**
Pivot: Mother Tongue vs Total Population
Included slicers for **Gender** and **Town**

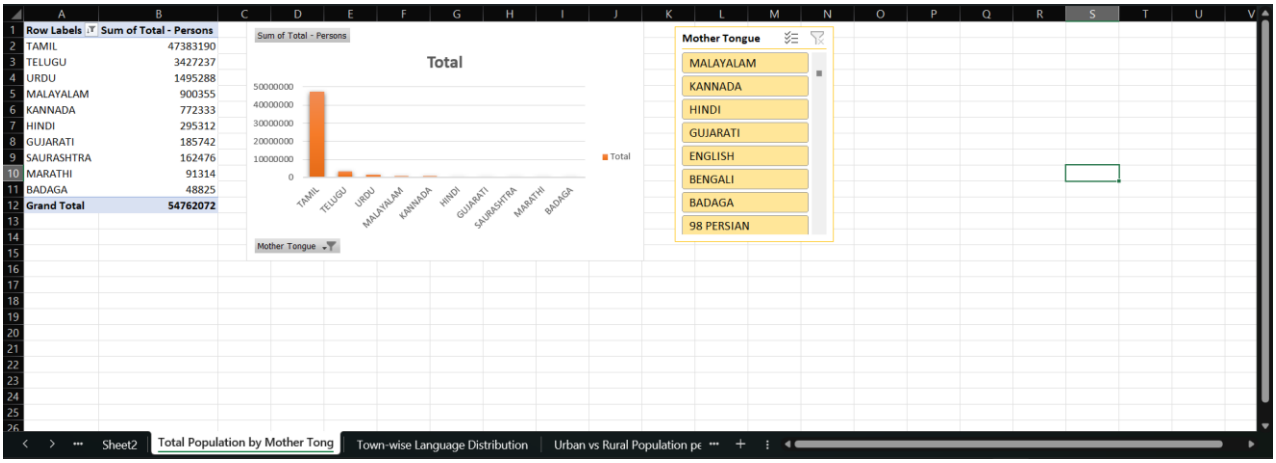
**iii. Functions Used:**

- PivotTable
- SUM aggregation
- Slicers
- Visual: **Bar Chart**

**iv. Result:**

- Tamil emerges as the most spoken language across all towns
- Other languages like Telugu and Urdu show regional concentrations

**v. Visualization:**
Bar with slicers for town and gender.




## Analysis 2: Gender-wise Language Distribution

**i. Introduction:**
Evaluate how language populations differ between males and females.

**ii. Description:**
Pivot: Mother Tongue vs Total Males and Total Females
Used 100% stacked bar chart

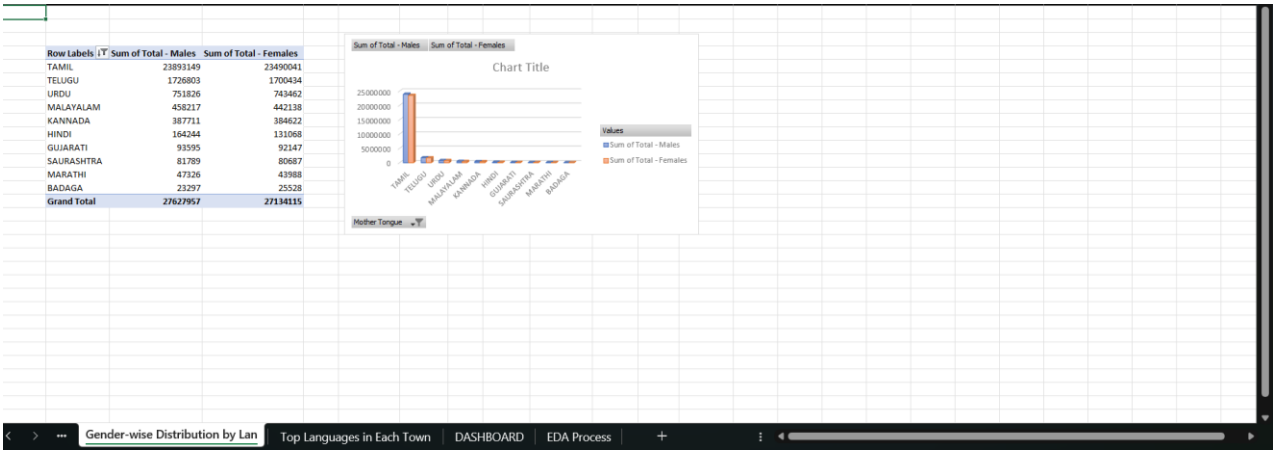
**iii. Functions Used:**

- `SUMIF()` to validate gender-based counts
- `ROUND()` for percentages
- Slicers: Town, Area Type

**iv. Result:**

- Gender distribution across languages remains consistent
- Slight skew observed in rural areas

**v. Visualization:**
100% Stacked Bar Chart




## Analysis 3: Urban vs Rural Language Speakers

### i. Introduction:

Understand how language usage changes in urban vs rural areas

### ii. Description:

Pivot: Mother Tongue vs Urban Population / Rural Population

### iii. Functions Used:

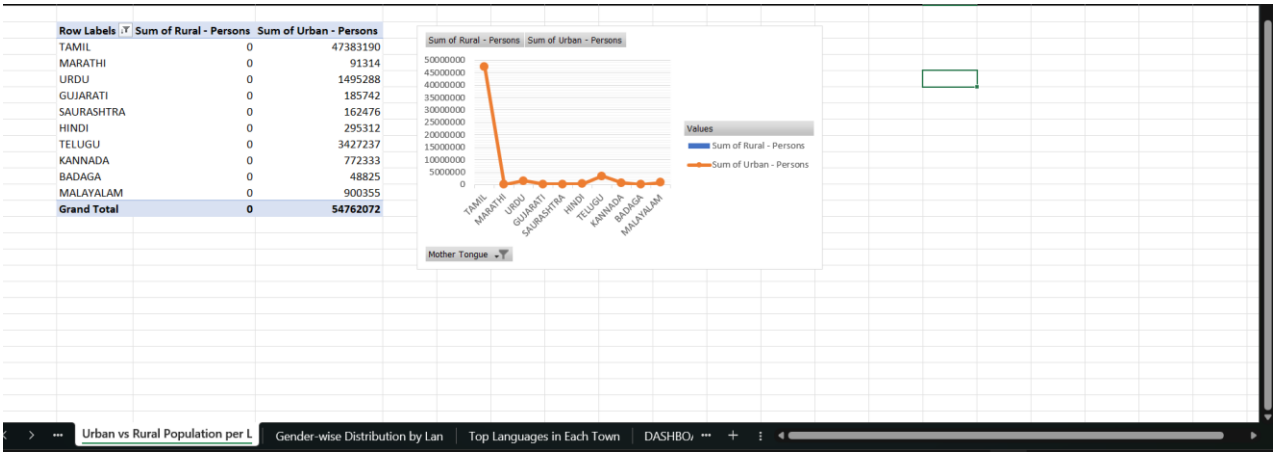
- Helper columns: Urban Share, Rural Share
- Chart: Line

### iv. Result:

- Urban areas show greater diversity
- Tamil is strong in both segments but minor languages more visible in urban towns

### v. Visualization:

Line with slicers for gender and town



## Analysis 4: Town-wise Comparison of Language Diversity

**i. Introduction:**
Compare number of distinct languages spoken in each town

**ii. Description:**
Pivot: Town vs Unique Mother Tongue Count
Used formula-based language counting

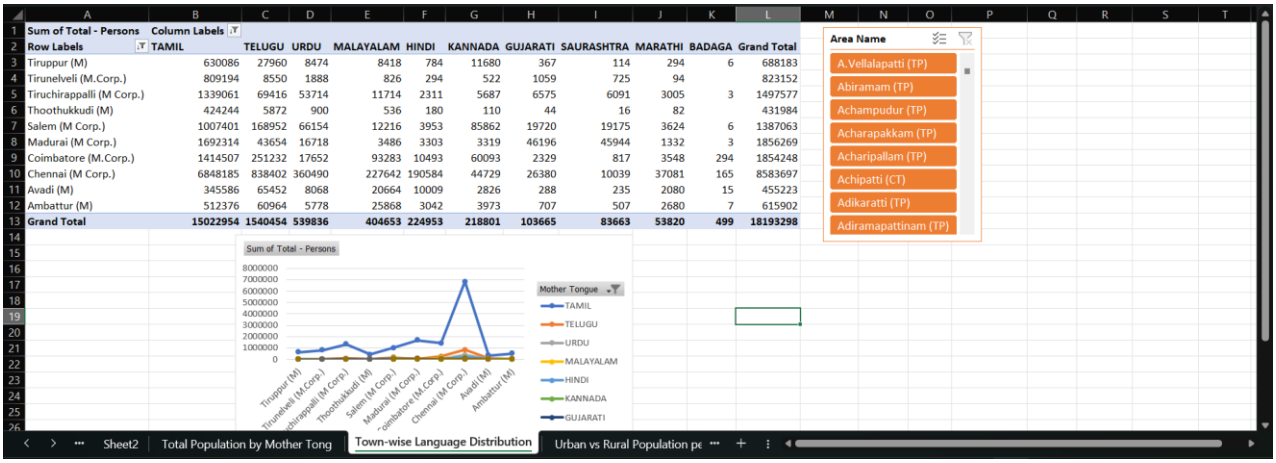
**iii. Functions Used:**

- `COUNTIFS()` for distinct combinations
- Line graph for circular comparison

**iv. Result:**

- Ponneri shows more language variety
- Arani has a concentrated linguistic profile

**v. Visualization:**
Line graph




## Analysis 5: Percentage Distribution of Languages

**i. Introduction:**
Visualize each language's share in its respective town population

**ii. Description:**
Calculated column: `% Share = Language Pop / Town Pop × 100`
Grouped by town

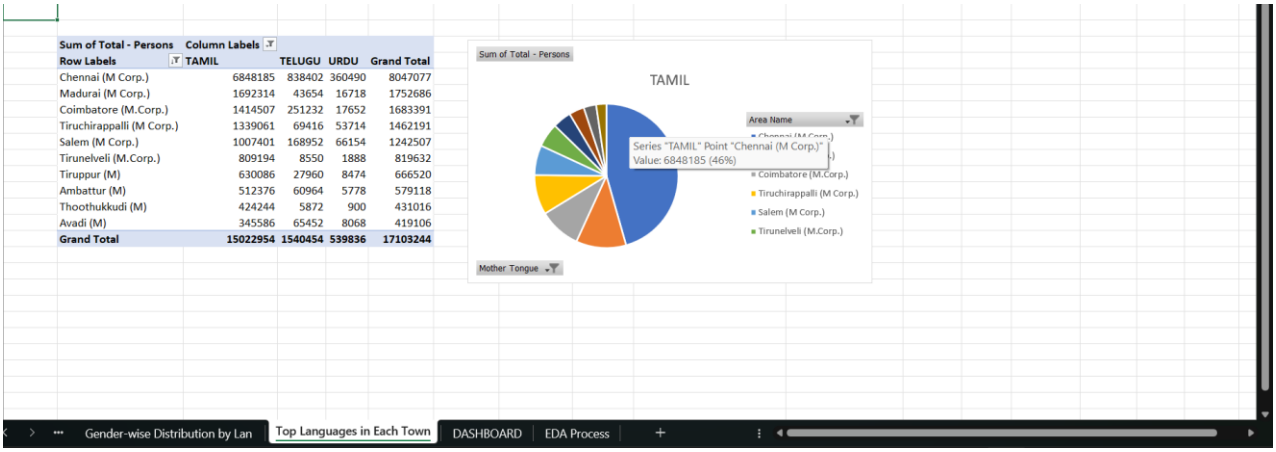
**iii. Functions Used:**

- `ROUND()`, `% Format`, `SUMIF()`
- Donut and Pie Charts

**iv. Result:**

- Tamil dominates with >80% in some towns
- Minority languages show strongholds in specific zones

**v. Visualization:**
Donut Chart per town, filterable by gender




# 5. Conclusion

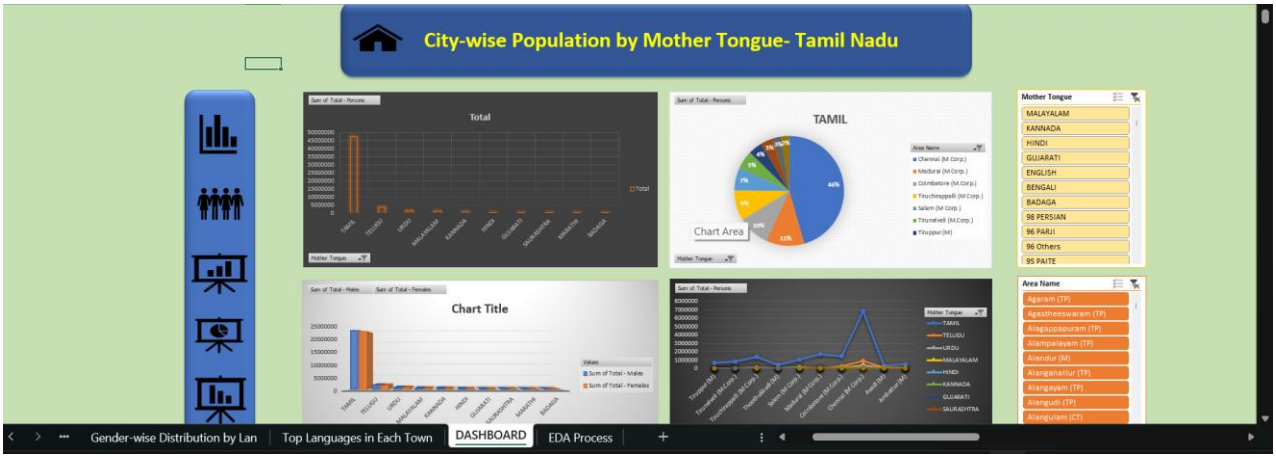
The project successfully leveraged **Excel** to conduct an in-depth **exploratory analysis** on Tamil Nadu's language distribution across selected towns.

**Key Takeaways:**

- Tamil is the dominant language in terms of volume
- Gender ratios are consistent across languages
- Urban towns tend to exhibit greater multilingual diversity
- Excel's PivotTables, slicers, and charts enable **interactive storytelling** with data

The insights from this project are valuable for:

- **Urban planners** designing multilingual public services
- **Education boards** preparing language-based curricula
- **Policymakers** considering minority language representation




# 6. Future Scope

- **Expand to more towns or full state** for comprehensive trends
- Add **age group distribution** to detect generational language shifts
- Apply **Power BI or Tableau** for interactive web dashboards
- Perform **geo-mapping** of language spread using Power Map or GIS
- Introduce **temporal analysis** (2001 vs 2011 census data)
- Implement **machine learning** to forecast future trends
- **Automate dashboards** using Excel Macros/VBA

# 7. References

1. Open Government Data (OGD) Platform India: data.gov.in

Link https://www.data.gov.in/resource/city-wise-population-mother-tongue-2001-tamil-nadu

2. LinkedIn: https://www.linkedin.com/posts/yuvanesh-mani-358994298_exceleda-dataanalysis-datavisualization-activity-7323059828900167682-eGLT?utm_source=share&utm_medium=member_android&rcm=ACoAAEgMYfQBW7Bi7OGoKSgtEZBe9rSFtZknApQ

1:04 3.39 KB/s 32%

**Yuvanesh Mani** · You
Attended Lovely Professional University
now ·

🗺️ City-wise Population by Mother Tongue – Tamil Nadu
📊 An Excel-Based Exploratory Data Analysis on Linguistic Demographics

I recently worked on a data analysis project using Microsoft Excel to explore how mother tongues are distributed across different cities in Tamil Nadu. The dataset, sourced from **data.gov.in**, offers valuable insights into the linguistic diversity of the state.

🔍 Key Insights:
City-Level Language Trends: Tamil dominates in most cities, but there's a noticeable presence of Telugu, Urdu, Hindi, and Malayalam speakers, especially in urban and border regions.

Population Patterns: Cities with more industrial and migrant populations show greater linguistic heterogeneity.

Cultural Implications: These findings reflect migration patterns, historical settlements, and urbanization trends.

📈 Excel Features Used:
Pivot tables for dynamic summarisation

Slicers for interactive filtering by language and city

Charts (bar & pie) to visually represent language distribution

💡 Takeaway:
This project highlights how even traditional tools like Excel can yield powerful insights when applied thoughtfully. Understanding language demographics helps policymakers, educators, and cultural researchers make more informed decisions.

#ExcelEDA #DataAnalysis #DataVisualization

Add a comment...

🔍 Key Insights:
City-Level Language Trends: Tamil dominates in most cities, but there's a noticeable presence of Telugu, Urdu, Hindi, and Malayalam speakers, especially in urban and border regions.

Population Patterns: Cities with more industrial and migrant populations show greater linguistic heterogeneity.

Cultural Implications: These findings reflect migration patterns, historical settlements, and urbanization trends.

📈 Excel Features Used:
Pivot tables for dynamic summarisation

Slicers for interactive filtering by language and city

Charts (bar & pie) to visually represent language distribution

💡 Takeaway:
This project highlights how even traditional tools like Excel can yield powerful insights when applied thoughtfully. Understanding language demographics helps policymakers, educators, and cultural researchers make more informed decisions.

**#ExcelEDA #DataAnalysis #DataVisualization #LinguisticDiversity #PublicData #DataGovIn #Demographics #PivotTables #CensusData #ExcelDashboard #CulturalInsights**