| **Project Title:** | **Census Data Standardization and Analysis Pipeline** |
| --- | --- |
| **Technologies** | **Python, SQL , MongoDB, Streamlit** |

**Dataset URL :** [Dataset](https://drive.google.com/drive/folders/10FLf8dEXqz_vc8p4DVoA5MKAh60gp1f6?usp=sharing)

**Problem Statement :**

The task is to clean, process, and analyze census data from a given source, including data renaming, missing data handling, state/UT name standardization, new state/UT formation handling, data storage, database connection, and querying. The goal is to ensure uniformity, accuracy, and accessibility of the census data for further analysis and visualization.

**Task 1: Rename the Column names**

For uniformity in the datasets and taking into consideration the census year, we need to rename some columns.

* State name to State/UT
* District name to District
* Male\_Literate to Literate\_Male
* Female\_Literate to Literate\_Female
* Rural\_Households to Households\_Rural
* Urban\_ Households to Households\_Urban
* Age\_Group\_0\_29 to Young\_and\_Adult
* Age\_Group\_30\_49 to Middle\_Aged
* Age\_Group\_50 to Senior\_Citizen
* Age not stated to Age\_Not\_Stated

**Task 2: Rename State/UT Names**

The State/UT names are in all caps in the census data, For uniformity across datasets we use the names so that only the first character of each word in the name is in upper case and the rest are in lower case. However, if the word is “and” then it should be all lowercase.

Examples:

* Andaman and Nicobar Islands
* Arunachal Pradesh
* Bihar

**Task 3: New State/UT formation**

* In 2014 Telangana was formed after it split from Andhra Pradesh, The districts that were included in Telangana are stored in *Data/Telangana.txt* . Read the text file and Rename the State/UT From “Andhra Pradesh” to “Telangana” for the given districts.
* In 2019 Ladakh was formed after it split from Jammu and Kashmir, which included the districts Leh and Kargil. Rename the State/UT From “Jammu and Kashmir” to “Ladakh” for the given districts.

**Task 4: Find and process Missing Data**

Find and store the percentage of data missing for each column.

Some data can be found and filled in by using information from other cells. Try to find the correct data by using information from other cells and filling it in. Find and store the percentage of data missing for each column.

Hint:

* Population = Male + Female
* Literate = Literate\_Male + Literate\_Female
* Population = Young\_and\_Adult+ Middle\_Aged + Senior\_Citizen + Age\_Not\_Stated
* Households = Households\_Rural + Households\_Urban

compares the amount of missing data before and after the data-filling process was done.

**Task 5: Save Data to MongoDB**

Save the processed data to mongoDB with a collection named “*census*” .

**Task 6: Database connection and data upload**

Data should be fetched from the mongoDB and to be uploaded to a relational database using python code . The table names should be the same as the file names without the extension.

The primary key and foreign key constraints should be included in the tables wherever required.

**Task 7: Run Query on the database and show output on streamlit**

1. What is the total population of each district?
2. How many literate males and females are there in each district?
3. What is the percentage of workers (both male and female) in each district?
4. How many households have access to LPG or PNG as a cooking fuel in each district?
5. What is the religious composition (Hindus, Muslims, Christians, etc.) of each district?
6. How many households have internet access in each district?
7. What is the educational attainment distribution (below primary, primary, middle, secondary, etc.) in each district?
8. How many households have access to various modes of transportation (bicycle, car, radio, television, etc.) in each district?
9. What is the condition of occupied census houses (dilapidated, with separate kitchen, with bathing facility, with latrine facility, etc.) in each district?
10. How is the household size distributed (1 person, 2 persons, 3-5 persons, etc.) in each district?
11. What is the total number of households in each state?
12. How many households have a latrine facility within the premises in each state?
13. What is the average household size in each state?
14. How many households are owned versus rented in each state?
15. What is the distribution of different types of latrine facilities (pit latrine, flush latrine, etc.) in each state?
16. How many households have access to drinking water sources near the premises in each state?
17. What is the average household income distribution in each state based on the power parity categories?
18. What is the percentage of married couples with different household sizes in each state?
19. How many households fall below the poverty line in each state based on the power parity categories?
20. What is the overall literacy rate (percentage of literate population) in each state?

**Submission:**

* Provide a well-commented Python file containing the complete code for the project, organized into sections for data Pipeline and Analysis.
* Upload the same into github with a proper Readme file.
* Presentation on the entire project, including Problem Statement, Tools Used, Approaches and Insights Found.

**Evaluation Metrics:**

* Project evaluation will be done in the live session and have to showcase the approaches done to complete the project
* You are supposed to write a code in a modular fashion (in functional blocks)
* Maintainable: It can be maintained, even as your codebase grows.
* Portable: It works the same in every environment (operating system)
* You have to maintain your code on GitHub.(Mandatory)
* You have to keep your GitHub repo public so that anyone can check yourcode.(Mandatory)
* Proper readme file you have to maintain for any project development(Mandatory)
* Follow the coding standards:
  + https://www.python.org/dev/peps/pep-0008/
* You should include basic workflow and execution of the entire project in the readme file on GitHub

**GitHub Repo:**

The attached reference document will help you use GitHub effectively. - [Link](https://docs.google.com/presentation/d/1XHCbgUOqbcXNUyQ87vTlKdKRgAbBxtkA/edit?usp=sharing&ouid=106590842700357786537&rtpof=true&sd=true)

**Reference Material:**

Official Documentation:

* <https://www.python.org/doc/>
* <https://docs.streamlit.io/>
* <https://dev.mysql.com/doc/>