**Data Engineering Capstone project on Census Data**

**Dataset url:** [**https://drive.google.com/drive/folders/10FLf8dEXqz\_vc8p4DVoA5MKAh60gp1f6**](https://drive.google.com/drive/folders/10FLf8dEXqz_vc8p4DVoA5MKAh60gp1f6)

**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

Colab url : [https://colab.research.google.com/drive/1ibym0wwvvB4hpH8Fw9dkkKjxhBXUBAka#](https://colab.research.google.com/drive/1ibym0wwvvB4hpH8Fw9dkkKjxhBXUBAka)

Tools Used:

Pandas, MongoDB, DataFrames, SqlAlchemy, Streamlit

Approaches:

1. Read the csv file using pandas
2. Data Pre-processing (Filling missing data, renaming column names)
3. Converted csv data into key-value pairs using to\_dict and column names into lower case for storing it in MongoDB
4. Used DataFrame Concept for storing MongoDB data into sql
5. Used Streamlit application for running the sql queries and displaying it in UI