**CAPSTONE PROJECT**

**CRIMES IN INDIA**

**C SOMA SEKHAR GOUD**

**BATCH.NO:DSG0123**

After a Quick Overview of all datasets.

1. The dataset consists of information on Indian crimes committed state & district wise.
2. District wise crimes committed against (IPC, SC, ST, Children, Women)
3. Dataset consists of data from a period of (2001-2012).

Since the dataset which was provided is from period of 2001-2012 we shall collect data, perform analysis and other steps in this period only.

Phase-1:

We don’t have any of population, literacy rate & Area of states data in the given dataset, so we extract the data from the online sources.

<https://censusindia.gov.in/>

<https://www.indiastat.com/specimen-tables/demographics>

<https://socialjustice.gov.in/common/76669>

literacy rate: <https://en.wikipedia.org/wiki/List_of_Indian_states_and_union_territories_by_literacy_rate>

After going through official websites it is found that survey is conducted for very 10 years(decade) so we get data for 2001 & 2011.

After analysing data, we can see that:

Top 5 States with Largest Area are Rajasthan, Madhya Pradesh, Maharashtra, Uttar Pradesh, Gujarat.

Top5 populated states both in 2001 & 2011 are Uttar Pradesh, Maharashtra, Bihar, West Bengal, Andhra Pradesh.

But when it comes to literacy rate there is a bit change in order and a state.

Top 5 Literate States in 2001: Kerala, Mizoram, Lakshadweep, Goa, Chandigarh

Top 5 Literate States in 2011: Kerala, Lakshadweep Mizoram, Goa, Tripura

**Phase-2:**

The dataset provided has state data along with district wish from 2001-2012

The dataset which we prepared in phase-1(df\_phase1) has literacy rate from 2001 & 2011 only. So, the feasible way is to use data from only those years.

**Phase-3:**

In this phase we are going to execute few sql queries and analyse data.

1. Database and Tables Creation:

Created a MySQL database named Capstone\_phase3.

And then we created three tables (against\_women\_2001\_2012, against\_ST\_2001\_2012, crimes\_committed\_IPC\_2001\_2012) of each cases to store data from different CSV files.

2. Data Insertion:

Using sqlalchemy we have Loaded data from the CSV file into MySQL

42\_District\_wise\_crimes\_committed\_against\_women\_2001\_2012.csv into the against\_women\_2001\_2012 table.

02\_District\_wise\_crimes\_committed\_against\_ST\_2001\_2012.csv into the against\_ST\_2001\_2012 table.

01\_District\_wise\_crimes\_committed\_IPC\_2001\_2012.csv into the crimes\_committed\_IPC\_2001\_2012 table.

3. **SQL Queries and Analysis:**

* We have Avoided where District column has TOTAL ,DELHI UT TOTAL as its irrelevant.

**3.1 - 3.2 - 3.3:**

**3.1:** Table creation

* we Identified the highest and lowest number of crimes (Rape, Kidnapping, Murders) in various districts and states.

MURSHIDABAD from West Bengal has highest number of rapes(568) & Kidnapping\_and\_Abduction(492).

After looking into Top10 we can see that West Bengal has most of the Rapes and Kidnapping\_and\_Abduction.

There are a lot of districts with ‘0’ rapes.

* Explored the distribution of these crimes and identified extreme values.

**3.4 - 3.5 - 3.6:**

Similarly Created a table against\_st\_2001\_2012 and pushed the respective data from .csv file into table.

* Focuses on specific crimes (Dacoity, Robbery, Murders) and identified the districts with the highest and lowest occurrences.
* Explored districts with 0 murders and identified the number of such districts.

Highest number of Dacoity(29) and Robbery(32)from DAHOD district.

A df\_36 DataFrame is created which consists all districts (810) with ‘0’ murders

**3.7:**

* Analyzed the number of murders in ascending order in district and year-wise.
* Identified trends and patterns in murder occurrences over the years.

**3.8.1 - 3.8.2 - 3.8.3:**

* Created a new table (**crimes\_committed\_IPC\_2001\_2012**) for specific crime data.
* Identified districts with the highest number of murders year-wise and analyzed the results.
* The query\_3\_8\_2 identifies, for each state and year, the district with the maximum number of murders using a subquery and left join in the crimes\_committed\_IPC\_2001\_2012 table, excluding 'TOTAL' districts.
* Stored the results in a DataFrame(df\_382) and we filtered districts that appear 3 or more than 3 years and stored as sorted\_df\_382 DataFrame.

**3.8.4:** We visualize thesorted\_df\_382 dataFrame using plotly, seaborn, matplotlib.