EAS 503-JLI - Python for Data Scientists - Project Report

Crime Trends and Incarceration Analysis

Members: Aarti Jadhav, Krish Shah and Samit Bhoir

Date: December 2024

---

1. Dataset Collection and Justification

The project aims to analyze crime trends and incarceration data to provide valuable insights into criminal behavior, juvenile crime trends, and the relationship between prison capacity and inmate populations.

Dataset Characteristics

The data includes crime statistics, categorized by different crimes such as **murder**, **rape**, **theft**, and **robbery**, along with demographic variables such as **age** and **gender**. The dataset also incorporates **prison statistics**, such as **inmate populations** and **facility capacities**, spanning over **15 years**.

Dataset Size:

- **Dimensions**: The dataset contains multiple categories, with data spanning 15 years.
- **Format**: The dataset was processed and visualized using tables, graphs, and scatter plots.

Key Characteristics:

- **Numerical Data**: Covers crime counts (e.g., number of murders, thefts) and demographics (e.g., age, gender).
- **Categorical Data**: Includes crime types and demographic information such as **age groups**, **gender**, **prison capacity**, and **inmate populations**.

Relevance:

- The dataset offers insights into **age-wise crime patterns**, helping to analyze how crime rates differ by age group.
- It highlights the **efficiency of the judicial system** and provides valuable data for **predictive analysis** in law enforcement and corrections planning.

---

2. Data Wrangling and Cleaning

Data cleaning was a crucial part of preparing the dataset for analysis. The following key steps were undertaken:

- **Missing Data Handling**: Identified and removed missing or inconsistent data points from key columns like **age**, **crime type**, and **gender**.
- **Standardization of Formats**: The numerical variables were normalized to ensure uniformity, making the data easier to analyze.
- **Encoding Categorical Variables**: Categorical variables like **crime type** and **gender** were encoded for statistical analysis, making them compatible with machine learning models.

These steps ensured that the data was cleaned, consistent, and ready for further analysis and visualization.

---

3. Visual Analysis and Observations

The project includes several visualizations to better understand the trends and patterns in the data: **Crime**

Data by Age Group

- **Observation**: The data shows that **murder** is the most dominant crime in the **18-30 age group**, followed by **rape** and **attempted murder**. **Theft** and **robbery** are significantly less frequent in comparison. This suggests that violent crimes are more prevalent in this age group than property crimes.

Youth Crime Trends (16-18 Age Group)

- **Observation**: The graph reveals a sharp **decline in youth crime** from **2001 to 2008**, with a slight uptick in **2013**. This could suggest improvements in law enforcement or judicial practices, or it could indicate broader social changes that reduced juvenile crime. **Prison Capacity vs. Inmate Population**
- **Observation**: A strong **correlation** was observed between **prison capacity** and **inmate populations**, both for **male** and **female** inmates. This highlights that **capacity constraints** play a major role in the total number of inmates housed, indicating that many facilities are nearing or operating at full capacity.

Interactive Crime Dashboard Overview:

The dashboard is built using **Streamlit** and provides a dynamic way to visualize crime trends across various categories over the years. Users can select crime categories (e.g., **Other Crimes**, **Personal/Violent Crimes**, **Property/Financial Crimes**) via a dropdown menu. The selected category is displayed as a **Plotly line chart**, showing the trend over the years. Additionally, separate charts for each crime type and the **Grand Total Trends** are available for deeper analysis. This allows for a flexible, user-driven exploration of the crime data. The **data overview** table further aids users in understanding the underlying dataset.

Fig 3.1

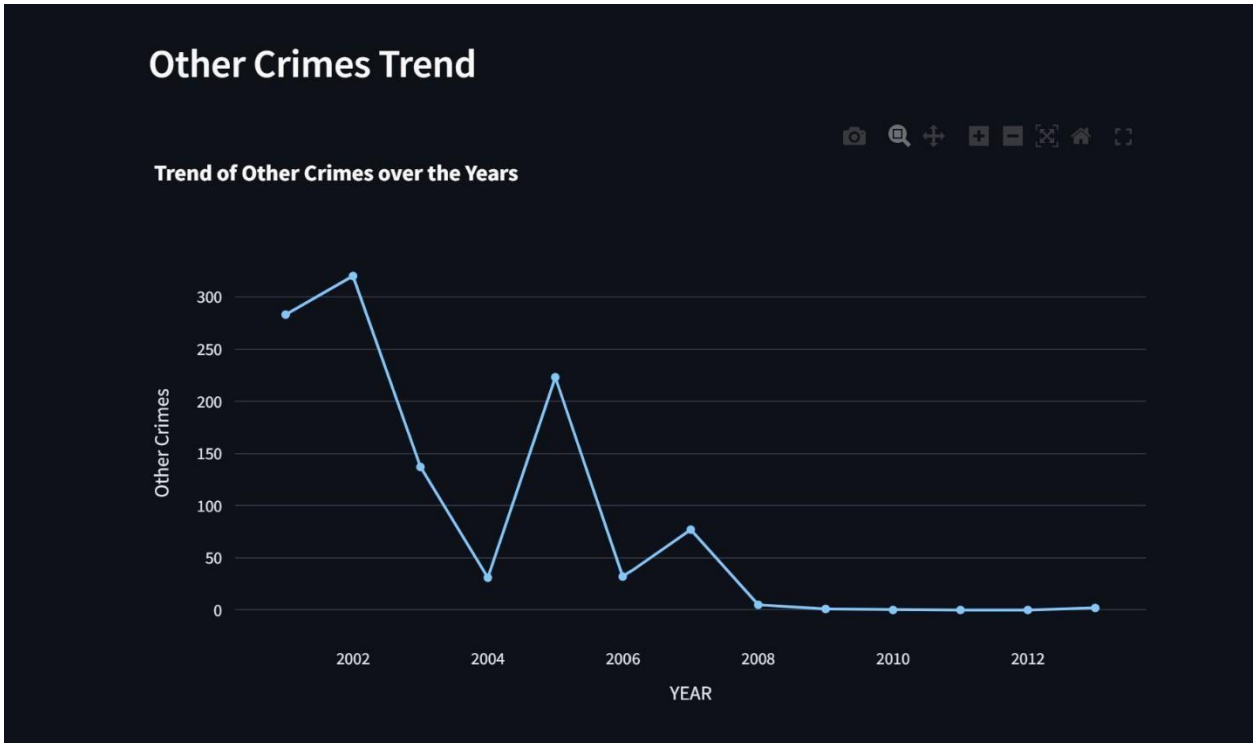

Other Crimes Trend
Trend of Other Crimes over the Years
300
250
200
150
100
50
0
Other Crimes
2002
2004
2006
2008
2010
2012
YEAR

Fig 3.2

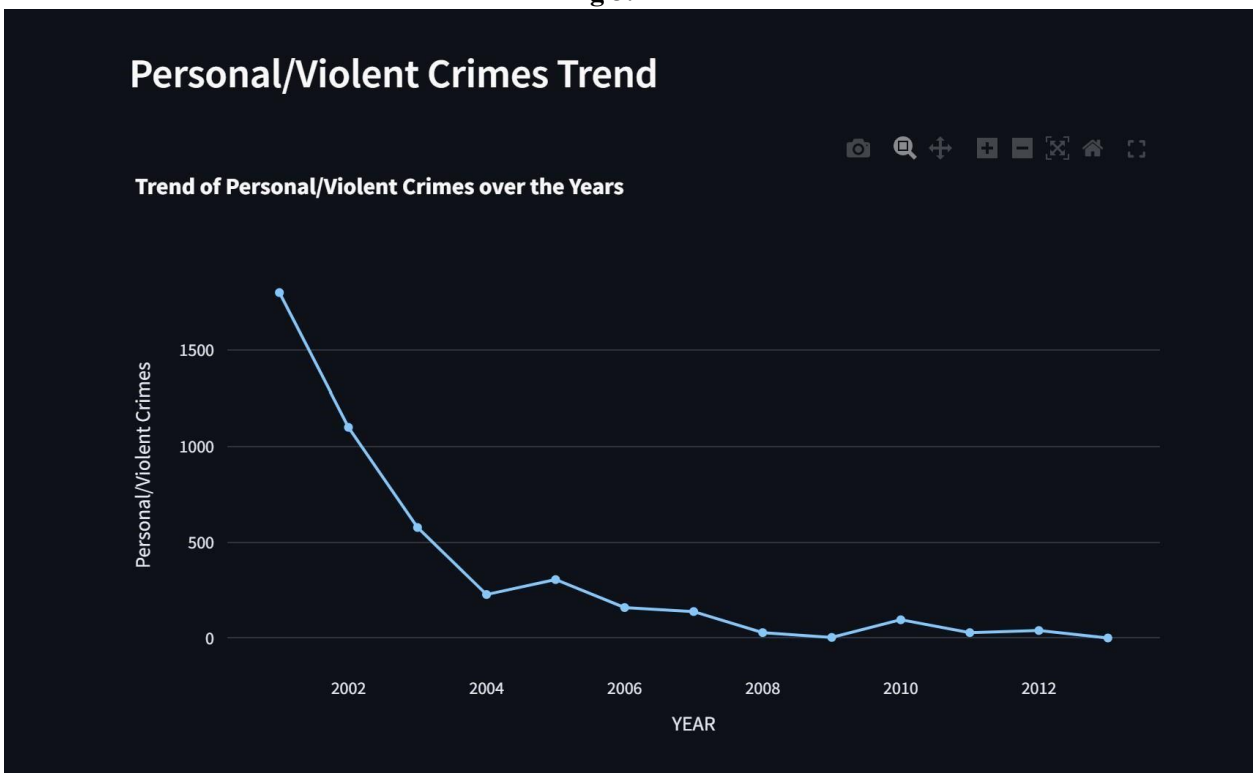

Personal/Violent Crimes Trend
Trend of Personal/Violent Crimes over the Years
1500
1000
500
0
Personal/Violent Crimes
2002
2004
2006
2008
2010
2012
YEAR

Fig 3.3

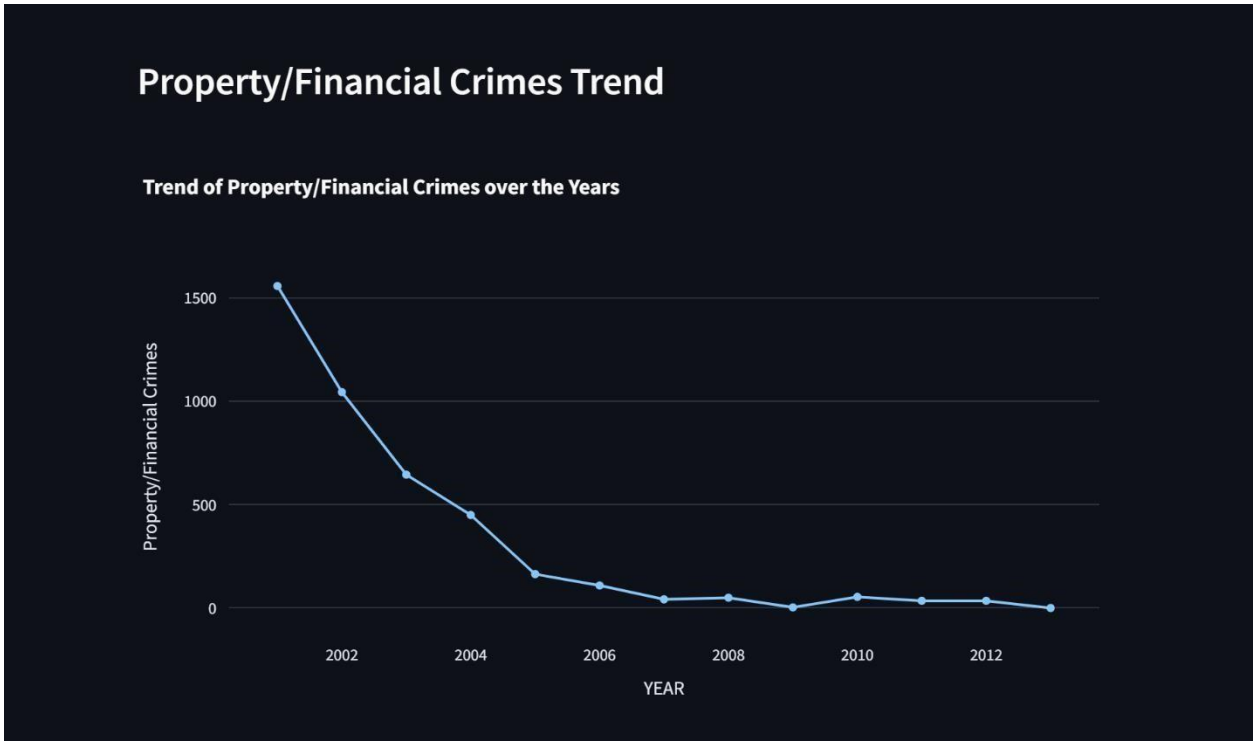


Fig 3.4

These visualizations provide essential insights into the relationship between crime trends, youth involvement, and the challenges faced by correctional institutions.

---

4. SQL Database Design and Analysis

To facilitate efficient querying and analysis, the dataset was structured into a **normalized SQL schema**:

- **Foreign Key Relationships**: These relationships were created between crime categories and demographic data, allowing us to perform more complex queries and analyses.
- The database design ensures that the dataset can be efficiently queried for deeper insights and can be extended for future analysis.

Table Name	Role	Description
States	Dimension	Stores unique names of states with **state_id** as PK.

AgeGroup	Fact	Stores age-wise distribution of inmates by state and year.
JailCapacity	Fact	Tracks capacity, population, and details of jails.
Escapees	Fact	Tracks prison escape details, including re-arrests.
IPC_Crime_UnderTrial	Fact	Details of inmates under trial for IPC crimes.
IPC_Crime_Convicted	Fact	Details of inmates convicted for IPC crimes.

Table	**Primary Key**	**Foreign Keys**
States	state_id	None
AgeGroup	agegroup_id	state_id (FK to States)
JailCapacity	jail_id	state_id (FK to States)
Escapees	escapee_id	state_id (FK to States)
IPC_Crime_UnderTrial	undertrial_id	state_id (FK to States)
IPC_Crime_Convicted	convicted_id	state_id (FK to States)

Normalization (1NF, 2NF, 3NF)

- **1NF**: All values are atomic, no repeating groups.

- **2NF**: Each table's non-key attributes depend only on the primary key. There are no partial dependencies because we use a single unique **Primary Key**
- **3NF**: No transitive dependencies exist.

Data Import Process

During this process we have loaded data from **CSV files** into the **6 tables** in the database. While most of the tables (**AgeGroup, JailCapacity, Escapees, IPC_Crime_UnderTrial, IPC_Crime_Convicted**) are directly populated using Oracle APEX's Data Load Wizard, the **States** table is populated using a **TempStates** table to ensure data integrity and uniqueness of state names.

The following CSV files were used to populate the database:

- **agegroup.csv**
- **population_and_capacity_of_jails.csv**
- **inmates_escapee.csv**
- **ipc_crime_inmates_under_trial.csv**
- **ipc_crime_inmates_convicted.csv**

Problem: The data for **state_name** is scattered across **5 CSV files**, not in a single file.

Solution: To avoid duplicating **state names** from multiple files, all unique **state names** from the 5 CSV files are loaded into a temporary table called **TempStates** which allows us for a **centralized point** to extract unique state names and avoid duplicates.

TempStates serves as a **staging table** to consolidate, clean, and filter the list of unique state names before populating the **States** table.

```sql
CREATE GLOBAL TEMPORARY TABLE TempStates (
    state_name VARCHAR2(255),
    is_state NUMBER(1)
) ON COMMIT PRESERVE ROWS;
```

Fig 4.1: SQL Query to create TempStates

Once all the CSV files are loaded into **TempStates**, **distinct state names** are extracted and inserted into the **States** table.

```sql
INSERT INTO States (state_name, is_state)
SELECT DISTINCT UPPER(TRIM(state_name)), is_state
FROM TempStates
WHERE state_name IS NOT NULL
  AND UPPER(TRIM(state_name)) NOT IN (SELECT UPPER(TRIM(state_name)) FROM States);
```

Fig 4.2: SQL Query to Insert Unique States Fact

and Dimension Tables

Why 6 Tables?

We created 6 tables because it provides better query performance and helps to normalize the data. These **6 tables** follows the **Star Schema** approach, which separates **Fact** tables from **Dimension** tables. **States (Dimension)**: Provides lookup for state names and IDs.

Fact Tables (5 Tables): Stores measurable and quantitative data.

AgeGroup: Tracks inmate age-group distribution.

JailCapacity: Tracks jail capacity for each state.

Escapees: Tracks escape details.

IPC_Crime_UnderTrial: Details for under-trial inmates**.**

IPC_Crime_Convicted: Details for convicted inmates.

Dimension Tables (1 Table):

States: Lookup table for state names, **state_id** is the primary key used as a foreign key in Fact Tables.

FACT TABLES 1.

AgeGroup:

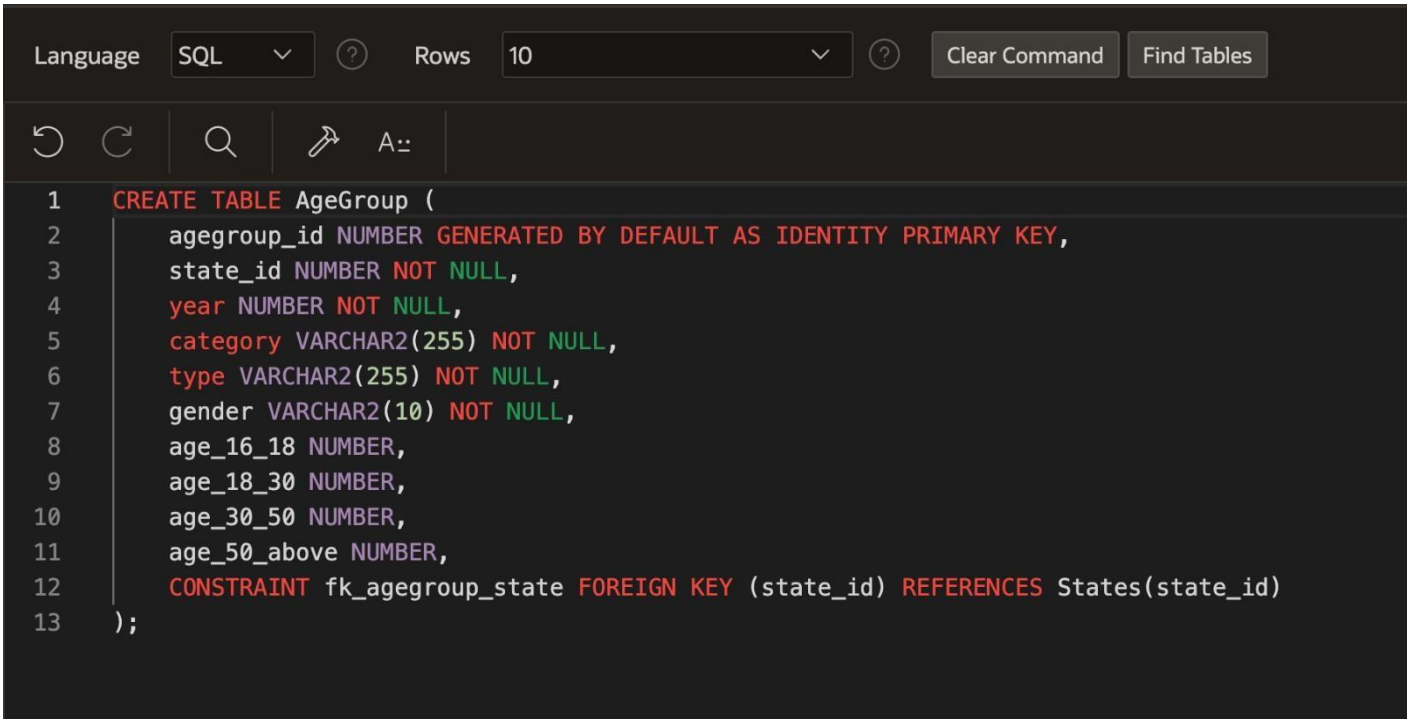


Fig 4.3: SQL Query to create AgeGroup

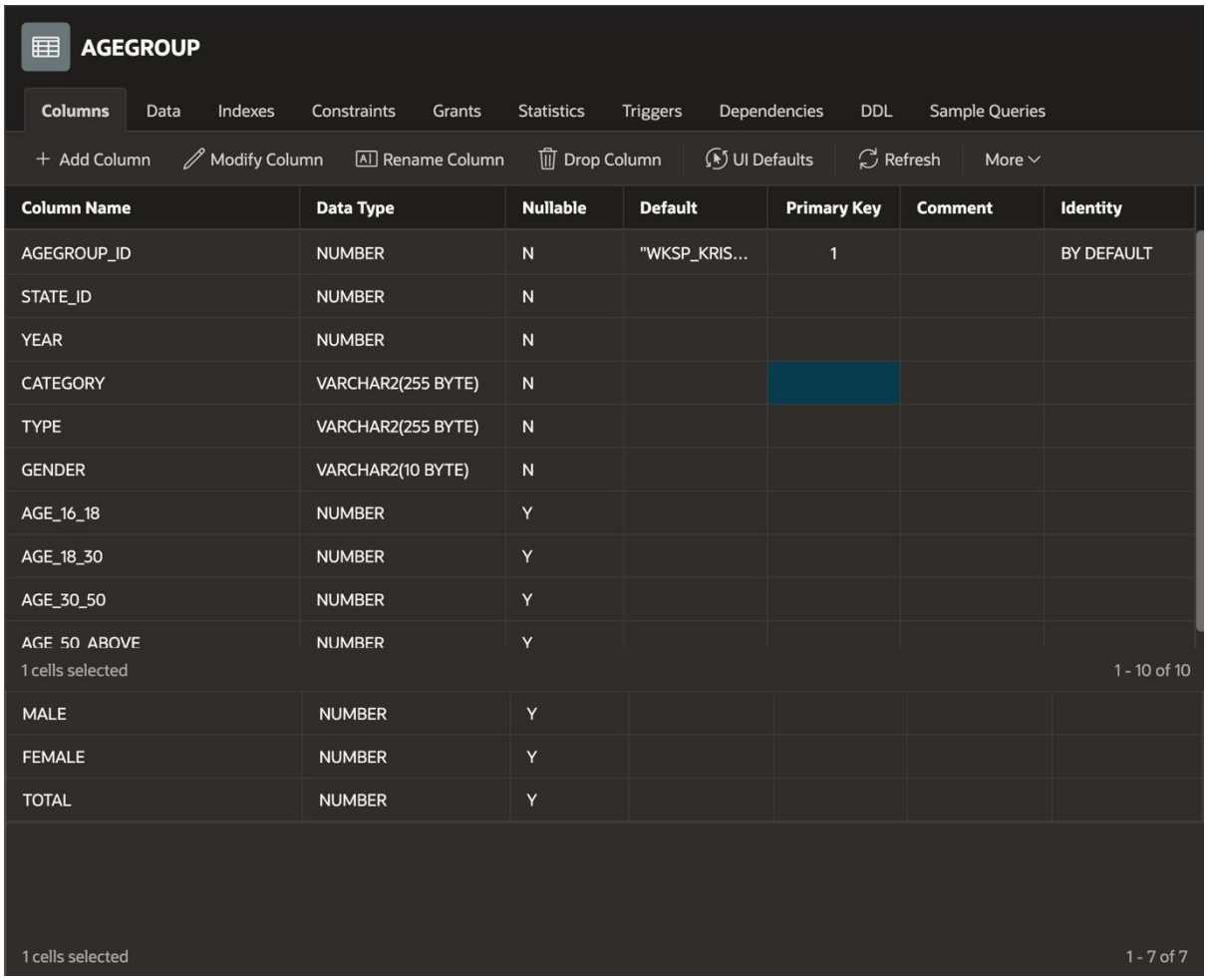
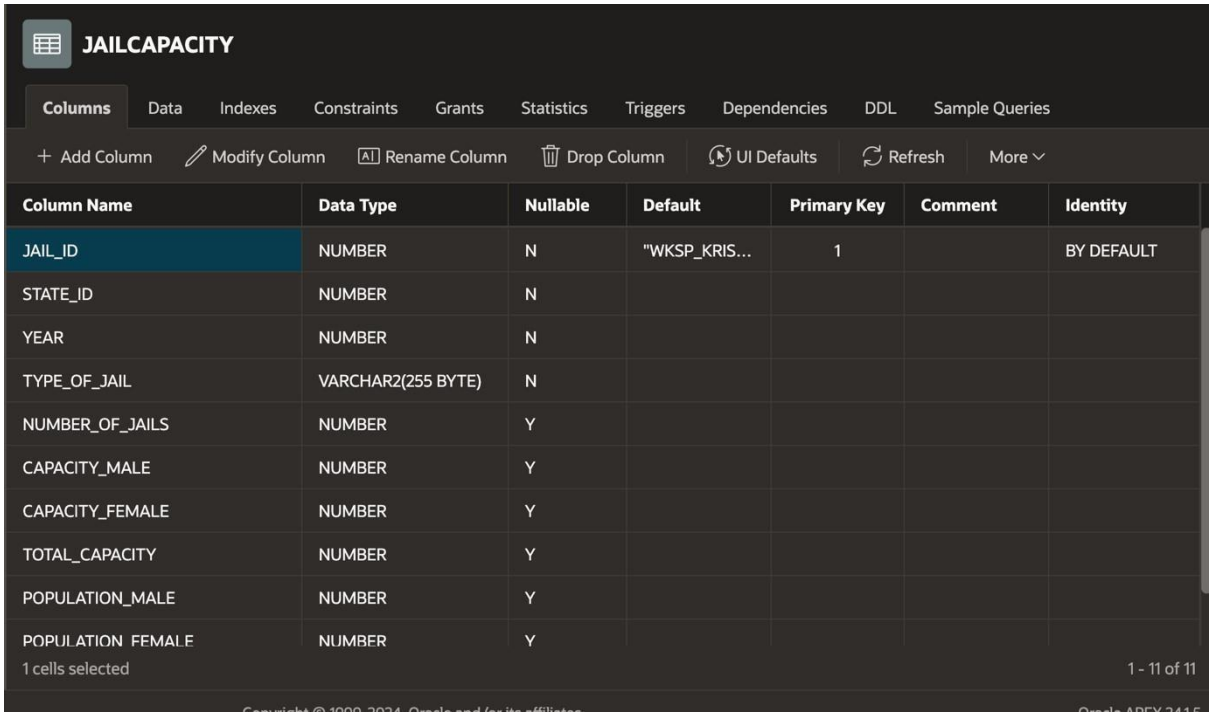


Fig 4.4: AgeGroup Table 2.

JailCapacity:

```
CREATE TABLE JailCapacity (
    jail_id NUMBER GENERATED BY DEFAULT AS IDENTITY PRIMARY KEY,
    state_id NUMBER NOT NULL,
    year NUMBER NOT NULL,
    type_of_jail VARCHAR2(255) NOT NULL,
    number_of_jails NUMBER,
    capacity_male NUMBER,
    capacity_female NUMBER,
    total_capacity NUMBER,
    population_male NUMBER,
    population_female NUMBER,
    total_population NUMBER,
    CONSTRAINT fk_jailcapacity_state FOREIGN KEY (state_id) REFERENCES States(state_id)
);
```

Fig 4.5: SQL Query to create JailCapacity



JAILCAPACITY

Columns | Data | Indexes | Constraints | Grants | Statistics | Triggers | Dependencies | DDL | Sample Queries

+ Add Column | Modify Column | Rename Column | Drop Column | UI Defaults | Refresh | More

Column Name	Data Type	Nullable	Default	Primary Key	Comment	Identity
JAIL_ID	NUMBER	N	"WKSP_KRIS...	1		BY DEFAULT
STATE_ID	NUMBER	N				
YEAR	NUMBER	N				
TYPE_OF_JAIL	VARCHAR2(255 BYTE)	N				
NUMBER_OF_JAILS	NUMBER	Y				
CAPACITY_MALE	NUMBER	Y				
CAPACITY_FEMALE	NUMBER	Y				
TOTAL_CAPACITY	NUMBER	Y				
POPULATION_MALE	NUMBER	Y				
POPULATION_FEMALE	NUMBER	Y				

1 cells selected

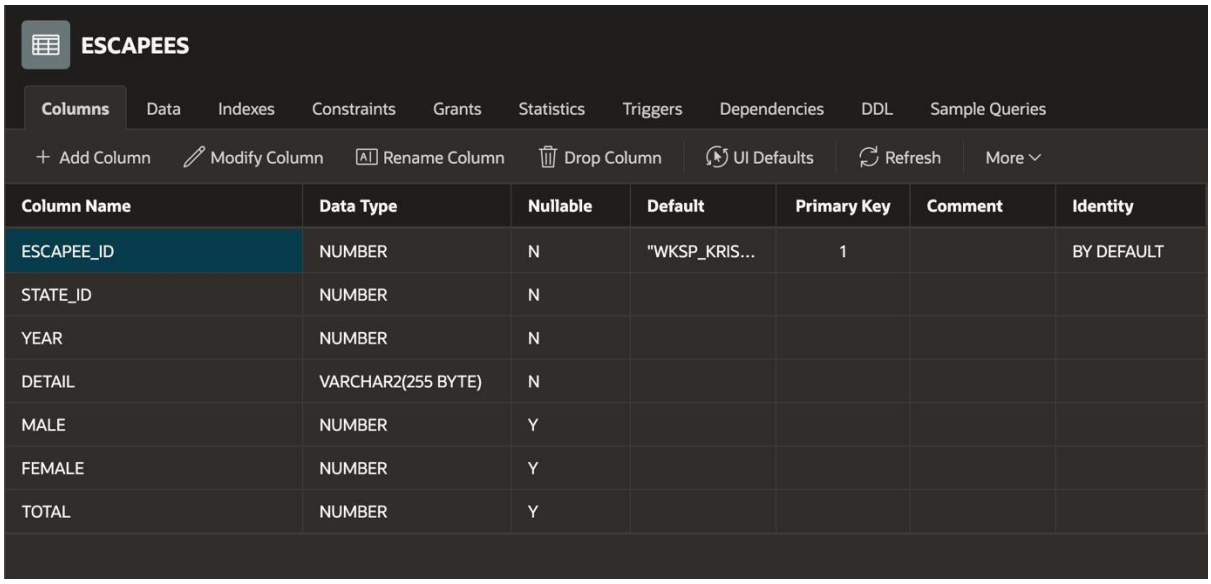
1 - 11 of 11

Fig 4.6: JailCapacity Table

3. **Escapees**:

```
1  CREATE TABLE Escapees (
2      escapee_id NUMBER GENERATED BY DEFAULT AS IDENTITY PRIMARY KEY,
3      state_id NUMBER NOT NULL,
4      year NUMBER NOT NULL,
5      detail VARCHAR2(255) NOT NULL,
6      male NUMBER,
7      female NUMBER,
8      total NUMBER,
9      CONSTRAINT fk_escapees_state FOREIGN KEY (state_id) REFERENCES States(state_id)
10 );
```

Fig 4.7: SQL Query to create Escapees



Column Name	Data Type	Nullable	Default	Primary Key	Comment	Identity
ESCAPEE_ID	NUMBER	N	"WKSP_KRIS...	1		BY DEFAULT
STATE_ID	NUMBER	N				
YEAR	NUMBER	N				
DETAIL	VARCHAR2(255 BYTE)	N				
MALE	NUMBER	Y				
FEMALE	NUMBER	Y				
TOTAL	NUMBER	Y				

Fig 4.8: Escapees Table 4.

IPC_Crime_UnderTrial:

```sql
CREATE TABLE IPC_Crime_UnderTrial (
    undertrial_id NUMBER GENERATED BY DEFAULT AS IDENTITY PRIMARY KEY,
    state_id NUMBER NOT NULL,
    year NUMBER NOT NULL,
    crime_head VARCHAR2(255) NOT NULL,
    male_16_18 NUMBER,
    female_16_18 NUMBER,
    total_16_18 NUMBER,
    male_18_30 NUMBER,
    female_18_30 NUMBER,
    total_18_30 NUMBER,
    male_30_50 NUMBER,
    female_30_50 NUMBER,
    total_30_50 NUMBER,
    male_above_50 NUMBER,
    female_above_50 NUMBER,
    total_above_50 NUMBER,
    total_male NUMBER,
    total_female NUMBER,
    grand_total NUMBER,
    CONSTRAINT fk_undertrial_state FOREIGN KEY (state_id) REFERENCES States(state_id)
);
```

Fig 4.9: SQL Query to create IPC_Crime_UnderTrial

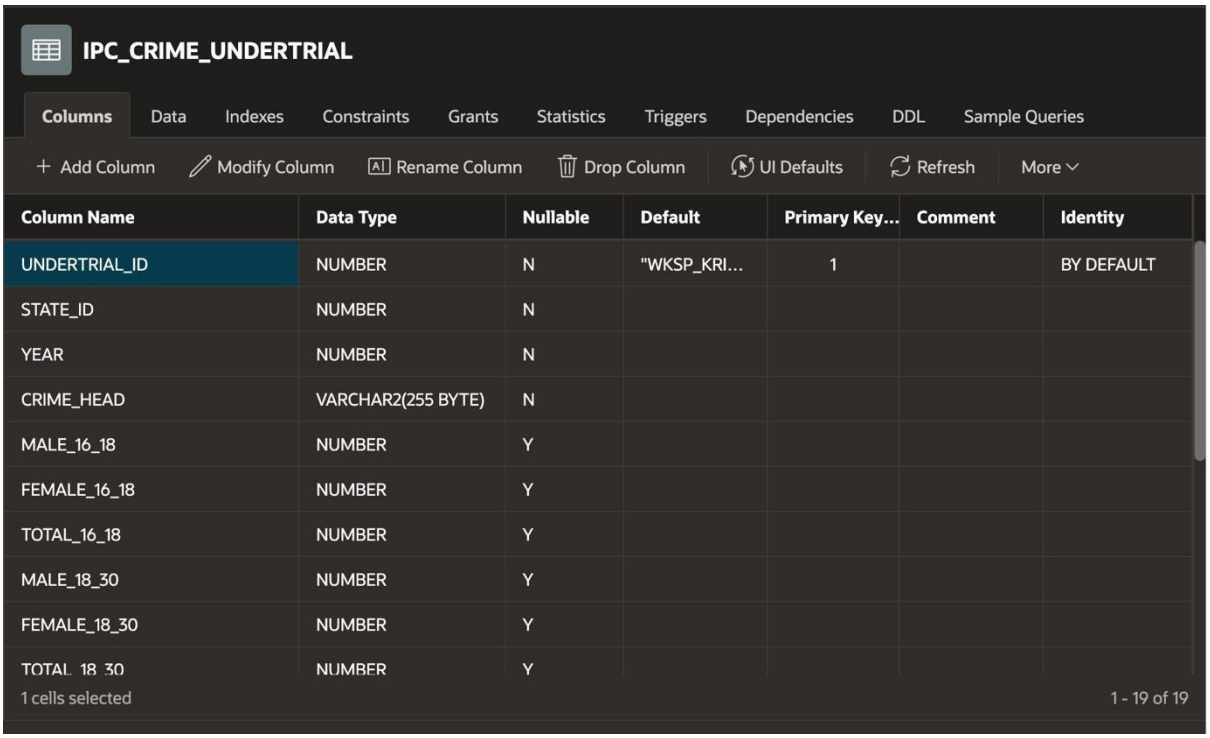


Fig 4.10: IPC_Crime_UnderTrial Table 5.

IPC_Crime_Convicted:

4.1.1

```sql
CREATE TABLE IPC_Crime_Convicted (
    convicted_id NUMBER GENERATED BY DEFAULT AS IDENTITY PRIMARY KEY,
    state_id NUMBER NOT NULL,
    year NUMBER NOT NULL,
    crime_head VARCHAR2(255) NOT NULL,
    male_16_18 NUMBER,
    female_16_18 NUMBER,
    total_16_18 NUMBER,
    male_18_30 NUMBER,
    female_18_30 NUMBER,
    total_18_30 NUMBER,
    male_30_50 NUMBER,
    female_30_50 NUMBER,
    total_30_50 NUMBER,
    male_above_50 NUMBER,
    female_above_50 NUMBER,
    total_above_50 NUMBER,
    total_male NUMBER,
    total_female NUMBER,
    grand_total NUMBER,
    CONSTRAINT fk_convicted_state FOREIGN KEY (state_id) REFERENCES States(state_id)
);
```

Fig 4.11: SQL Query to create IPC_Crime_Convicted

IPC_CRIME_CONVICTED

Columns | Data | Indexes | Constraints | Grants | Statistics | Triggers | Dependencies | DDL | Sample Queries

+ Add Column | Modify Column | Rename Column | Drop Column | UI Defaults | Refresh | More

Column Name	Data Type	Nullable	Default	Primary Key	Comment	Identity
CONVICTED_ID	NUMBER	N	"WKSP_KRIS...	1		BY DEFAULT
STATE_ID	NUMBER	N				
YEAR	NUMBER	N				
CRIME_HEAD	VARCHAR2(255 BYTE)	N				
MALE_16_18	NUMBER	Y				
FEMALE_16_18	NUMBER	Y				
TOTAL_16_18	NUMBER	Y				
MALE_18_30	NUMBER	Y				
FEMALE_18_30	NUMBER	Y				
TOTAL_18_30	NUMBER	Y				

1 cells selected

1 - 19 of 19

Fig 4.12: IPC_Crime_Convicted Table

DIMENSION TABLES

States:

```
1  CREATE TABLE States (
2      state_id NUMBER GENERATED BY DEFAULT AS IDENTITY PRIMARY KEY,
3      state_name VARCHAR2(255) UNIQUE NOT NULL,
4      is_state NUMBER(1) NOT NULL
5  );
6
```

Fig 4.13: SQL Query to create States

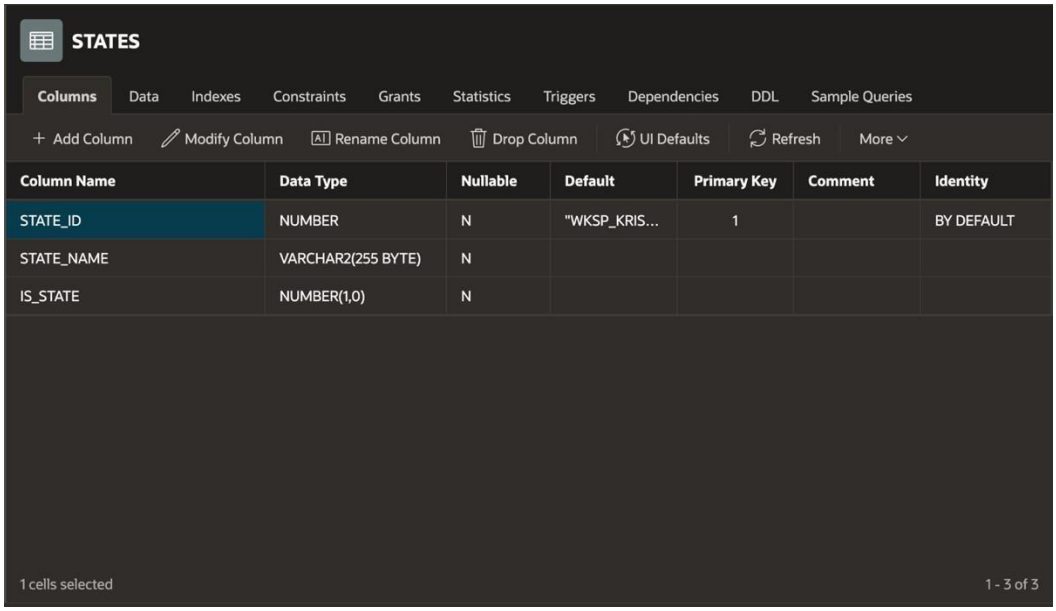
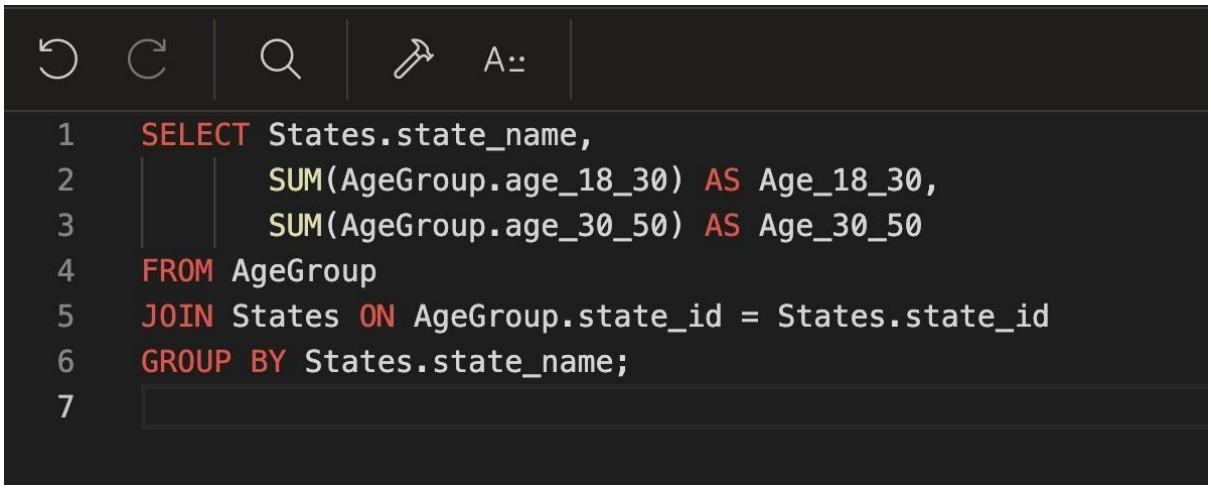


Fig 4.14: States Table

SQL QUERIES AND ANALYSIS

Here are a few SQL examples:

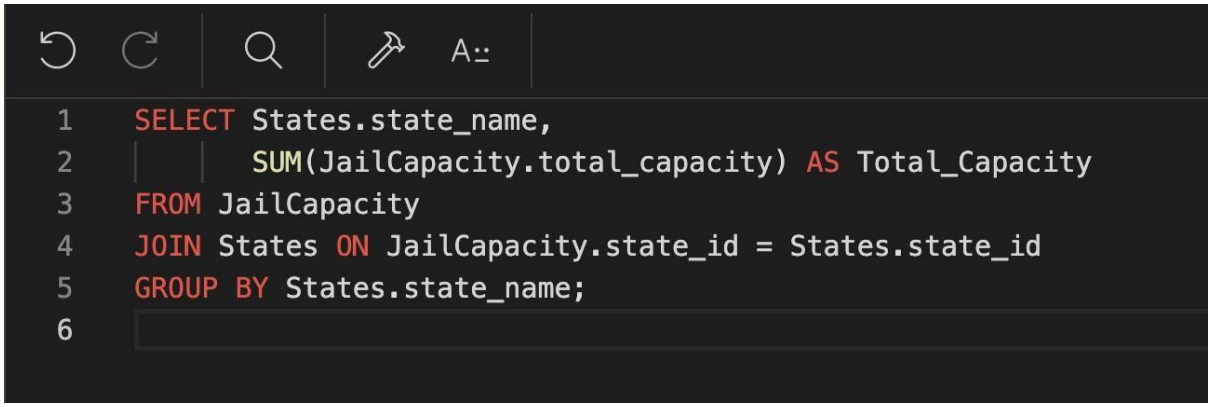
Query 1: Count of Inmates by State and Age Group



```
SELECT States.state_name,
       SUM(AgeGroup.age_18_30) AS Age_18_30,
       SUM(AgeGroup.age_30_50) AS Age_30_50
FROM AgeGroup
JOIN States ON AgeGroup.state_id = States.state_id
GROUP BY States.state_name;
```

Fig 4.15: SQL Query

Query 2: Total Jail Capacity per State



```sql
SELECT States.state_name,
       SUM(JailCapacity.total_capacity) AS Total_Capacity
FROM JailCapacity
JOIN States ON JailCapacity.state_id = States.state_id
GROUP BY States.state_name;
```

Fig 4.15: SQL Query

---

5. Conclusions and Future Work

This project has provided valuable insights into crime and incarceration trends. The key findings include:

- **Demographic insights** revealing that **violent crime** is more prevalent among younger age groups.
- A **strong correlation** between prison capacity and the total inmate population, emphasizing the challenges of overcrowding in prisons.

Future Work:

- **Predictive Modeling**: Incorporating machine learning techniques to model future crime trends and predict factors influencing incarceration rates.
- **Policy Impact Evaluation**: Exploring the effects of different law enforcement and correctional policies on crime rates and prison populations.

By extending this analysis, future work could provide even deeper insights into the factors that influence crime rates and the operation of correctional facilities, helping policymakers improve crime prevention and correctional management.