# ETL Application Report

# By Genti KOSUMI

**Police fatalities (Data World) and US** [**Crime**](https://datawrapper.dwcdn.net/vXnEw/4/) **rates by County (Kaggle)**

**Extraction:**

* Extracted data from two different csv files (sources – last page of this document) using pandas read\_csv
* <https://data.world/awram/us-police-involved-fatalities>
* <https://www.kaggle.com/mikejohnsonjr/united-states-crime-rates-by-county>

**Transformation:**

**Police Fatalities -** *Individuals killed by police in the US from 2000 – 2016*

This dataset aims to provide insight into individuals who were killed during altercations with police. It includes information on their age, race, mental health status, weapons they were armed with, and if they were fleeing.

The purpose of transforming this dataset was to group the data by state and by city, and to group them by Gender and by Race to have a clear view.

At the end we have 2 data frames, the first one has the police fatalities grouped by state and the second one is the one which has the police fatalities grouped by cities.

## **Crime rate by county -** *County-level crime data of the United States*

This dataset is about crime rate in the counties in the USA. The first thing I had to do was to split the “county\_name” column into 2 columns (county and state).

The purpose of transforming this dataset was also to group the data by state. Transforming this dataset was easier than the first one. After splitting the “county\_name”, I grouped the data by state.

**Loading**

* I used the elephantSQL because I couldn’t access postgres.
* I created the database connection to elephantSQL using SQLAlchemy.
* I loaded 3 tables in database named as “State”, “Fatalities” and “Cities”.

**Purpose**

The reason I have chosen these datasets was to see if there is any correlation between the crime rate with the number of fatalities caused by police in specific city and state.