**Advanced Databases/Databases Technologies 2022/2023 Project Report**

By Inês Fragoso 54454,

Robert Morosan 54456,

Milosz Wloch 59437,

Martim Silva 51304.

**Source of the dataset:** <https://www.kaggle.com/datasets/kwullum/fatal-police-shootings-in-the-us?select=PoliceKillingsUS.csv>

**Description of how to replicate the project creation:**

1. sqlite\_create\_db, mongodb\_create\_db -> creation of the databases
2. sqlite\_queries, mongodb\_queries -> running the queries, optimization and indexing for sqlite/mongodb

**Description of the dataset**

This dataset is a compilation done by the American newspaper "The Washington Post" since the start of 2015 of some of all of the US police shootings that resulted in casualties. The information gathered includes the race, age, gender of the person who was fatally shot, if the person was carrying a weapon or not, when/where this shooting took place, whether the victim showed signs of mental illness, if the police officer had a camera on their person at the time, among other pertinent information about each shooting. This dataset also contains information regarding the median household income per each city; the percentage of people over the age of 25 that have completed high school per each city; the poverty rate per each city and the shares of people of a given race (from white, black, Hispanic, Asian, and native American) per each city. All of this information was taken from the United States Census.

**Scheme used for both databases:**

Diagram, schematic

Description automatically generatedIt is worth noting that in this project, it was decided to keep the same database schemas for both types of infrastructures, the diagram used to represent them is based on the Unified Modelling Language standard and is presented below.

Note: For SQLite implementation the PoliceKillings id field is auto incremental

**Discussion of point done/not done in the project:**

1. For the first goal of the project, we selected a dataset in the Kaggle website that was the "Fatal Police Shootings in the US" by Karolina Wullum. We decided to select this dataset over others as it provides a big quantity of relatively clean data, more specifically a small amount of missing/mismatched data also this topic provides interesting information upon its further exploratory analysis. It is worth mentioning that every CSV file must have at least one column that are somehow linked to one another (this allows the setting of foreign keys when converting these files to collections). Anyway, before we could make use of this data by implementing a database, we had to create some constraints as primary and foreign keys to relate them. A UML diagram was created to present attributes and relations between tables. Some additional tables have been added to enable better data maintenance in the future.
2. For the second goal of the project, we were expected to create the database for the relational infrastructure in SQLite and the NoSQL infrastructure in MongoDB, while maintaining the overall structure of the data itself set in the diagram that was previously built. To be able to import the CSV files into databases we had to do some data pre-processing using Pandas that was, for example, changing the values of the various races into a single format across all CSVs among others. Furthermore, we removed ambiguous portions of the city names (namely "town", "city", "CDP") that would result in some discrepancies when connecting the data to one another.

CONTINUE THIS GOAL

1. #TODO
2. #TODO