Fatal Police Shootings in the USA

Andrei Martins, Daniel Dedoukh, Hidayat Rzayev, Ivaylo Gatev, Spyridon Ntokos
University of Applied Sciences Ulm
martins@mail.hs-ulm.de,
dedoukh@mail.hs-ulm.de,
ryazev@mail.hs-ulm.de,
gatev@mail.hs-ulm.de,
ntokos@mail.hs-ulm.de

***Abstract*—Our project is based on CRISP-DM Model. After finding the proper datasets for our project, we went to the Business Understanding phase, where we defined the goals and critetia for our project. The main goal is to show how different aspects are related to the killings by the USA police. In Data Understanding and Data Preparation phases we inspected the data and applied Data Cleansing. After all the data has been prepared and harmonized, we created the CDWH with the dimensions and a fact table. For analytical purposes, we created a Data Mart out of our CDWH and performed the analysis. The key findings of the analysis are the amount of killings based on race, weapon, time, location and age.**

I. Introduction

A. Scenario

Our goal in this project report is to show how social conditions as race, gender, income and others are related to the killings caused by the police in the USA. By analysing the data, we found patterns on how these killings are distributed among the population.

To achieve our objective of getting insightful meaning from the data, we will follow the *CRISP-DM process* [1]. Our main resources are 7 csv files that contain the data to be analysed. This data will be harmonized through the Staging Area before being inserted into the Database. The main tools used during the process where phpMyAdmin, MYSQL Workbench and RStudio.

B. Structure of the Paper

The paper is structured as follows: in Section II the process of extraction of the datasets is described. Then, in Section III the Staging Area and the data harmonization is detailed. Section IV describes how the Database was created, its dimensions and how they were combined for the Data Mart. In Section V we show the relationship between the different data and some of graphics that represent these relationships visually. Finally, we conclude our work in Section VI with the result of our analysis and what meaningful information we could get from it.

II. Open Data: Police Shootings in the US

As part of the *Data Understanding* phase for this data science project, the two available data sources are described: The CSV file *PoliceKillingsUS*, taken from [3] contains the shootings that happened between years 2015 to 2017. There are four additional CSV files from that data source, which contain US census data on poverty rate, high school graduation rate, median household income, and racial demographics. The second CSV file, *Police Fatalities* [2] contains data on fatalities dating back to 2000 up to year 2016.

III. Data Preparation

The first step of the data preparation phase is extracting the data from our initial data sources and importing it into 2 MySQL databases, one for each datasets. This is a straightforward process, as we had to write one LOAD DATA statement for each of the .csv files, with the correct field and line termination symbols. For each field, we assigned an appropriate data type, as well as an appropriate primary key for each table.

1) Police Fatalities Staging Area - The main tables(containing information about the police fatalities) from our 2 datasets have a similar structure, with only a slight difference in the columns. We selected only the fields, which are relevant for our analysis and are available in both of the datasets. But despite all of this, the representation of some of the data is a little different. For example the values for race and gender are represented with their full names in one of the datasets, and with only the first letter in the other. Our strategy

of dealing with this problem is deciding on one particular format for each field and storing the data in this unified way in the staging area. To avoid lower/upper case conflicts, we saved string values in upper case letters only. There are also some null values, which were automatically replaced by the LOAD DATA statement with an empty string for the VARCHAR type or 0 for numeric types. We stored the empty strings as 'UNKNOWN' in the staging area and kept the 0 as our value, which represents the unknown state, for the numeric types.

To harmonize the datasets we first stored the data for each one into a temporary table, applying the transformations described previously. However, some duplicate values existed in the bigger dataset, as well as matching values between both of them. The problem is that not all of the duplicates are exact matches. For example 2 records contain the same values for all but one column, in particular for the columns race, weapon and age. One record would contain a concrete value for one particular field and the other one an unknown value. Naturally, we kept the record that contains more information about the fatality.

Mostly there were only 2 duplicate rows for one fatality, but there were a few cases with 3 duplicates, which we dealt with manually. After that, we deleted the records that match ones from the smaller dataset. In the end, the temporary tables were merged into one. To make the ids more consistent, we dropped the original id columns and created our own auto increment primary key. Later, while working on the location dimension, we noticed that there was a typo in one of the records with the city Albuquerque and fixed it manually.

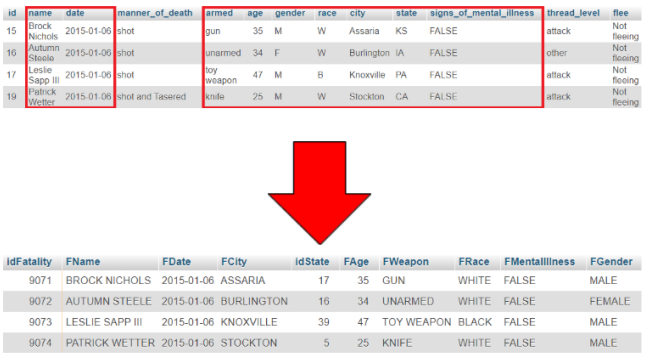

Fig. 1. Overview of solution strategy

2) Age and Weapon Maping Area - The Age Maping consist of the table MapAgeGroup. The age groups were defined by an interval with an appropriate lower and upper bound(age group 'UNKNOWN' for 0 values).
For the Weapon Dimension, there are 70 distinct weapons used in our staging area 's 'Fatality' table. Using a "CASE WHEN ... THEN ..." SQL syntax, we 've classified them in 10 distinct categories. This mapping table was stored in an intermediate database together with the Age maping tables.

IV. Concept for CDWH

The CDWH approach is visualized in Figure 2.

1) Time Dimension - The script used for the time dimension tables is a simplified version of the script written by Prof. Dr. Markus Goldstein. A procedure is used to build the calendar for days, months and year.
2) MentalIllnes dimension - A True / False value is used to determine the mental state of the victim.
3) Race dimension - Describes which race the victim belonged to. The values are: white, black, hispanic, native, asian, other or unknown.
4) Gender dimension - Describes the gender of the victim: male, female or unknown.
5) Weapon Dimension - Consisted of two tables, namely "Weapon" and "WeaponClass" which were populated according to the weapon mapping table from an intermediate database, with auto-increment IDs. Unknown weapon is declared as "UNKNOWN".
6) Age Dimension - The Age Dimension consists of the tables Age and Age Group. The age groups from the mapping table, without the lower and upper bound, were copied into the CDWH Age Group table. For the Age table, all of the distinct age values were taken from the staging area, mapped to their appropriate age group and inserted into the Age table of the CDWH. The age values themselves are used as a primary key to the table.
7) Location Dimension - consists of the tables City and State. Since the states were available in the Staging Area, they were simply copied from there. The distinct cities were taken from the Fatalities table of the Staging Area, as well as various statistics for each city, such as poverty level, median household income, etc. were taken from one of the initial data sources. Cities, for which statistics was not available, are denoted with 0 values.

For the creation of the DM, 4 tables were kept exactly as the same from the CDWH: Gender, MentalIllness,

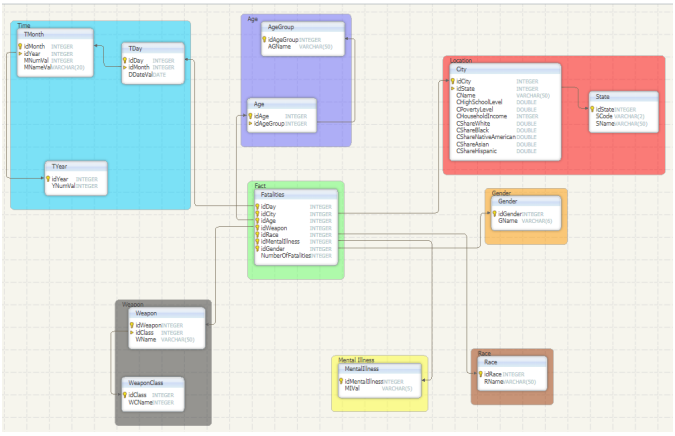

Fig. 2. Overview of the CDWH

Race and Fatalities (our fact table). Each of these tables were already alone in the dimensions of the CDWH, so there was no need to aggregate their information. The Fatalities table, which already had its data aggregated in the CDWH from 2 different tables in the Staging Area was also already prepared for the DM. The 4 remaining dimensions were aggregated as follows:

1) Time dimension: TYear, TMonth and TDay were aggregated into Tday.
2) Age dimension: AgeGroup and Age were aggregated into Age.
3) Weapon dimension: Weapon and WeaponClass were aggregated into Weapon.
4) Location dimension: City and State were aggregated into City.

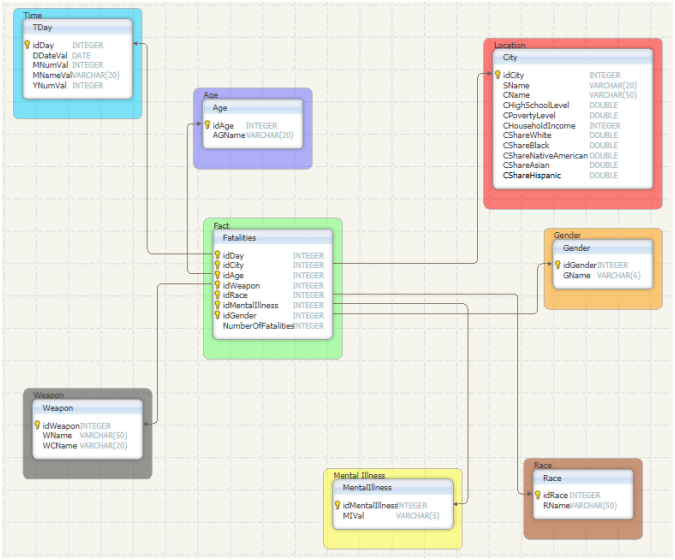

Fig. 3. Overview of DM

V. Analytics

Using our DM, we have created a view consisting of our fact table joined with all the other dimensions, thus creating a high-granularity table, which we exported as a CSV file and produced the following analysis plots using RStudio.

Data visualization:

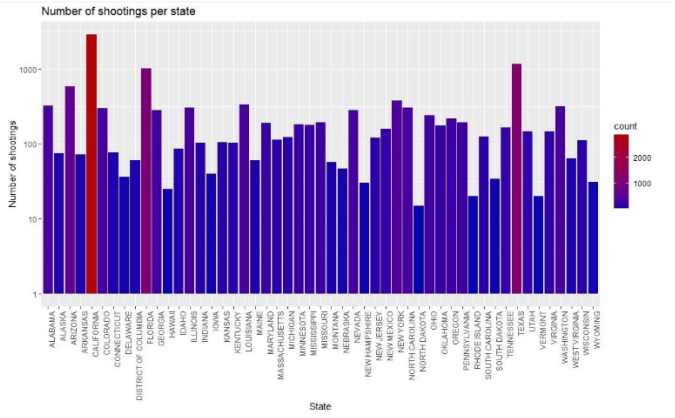

Fig. 4. It is clearly shown that the states with the highest number of police fatality shootings are California, Texas, Florida, etc.

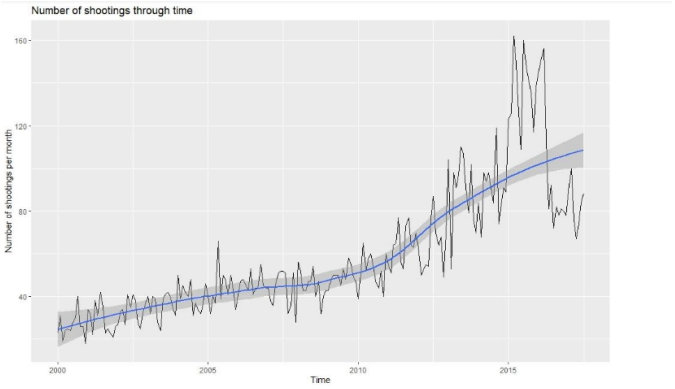

Fig. 5. Monthly reported fatal police shootings have increased through time, with the biggest spike reported shortly after the 2016 presidential election.

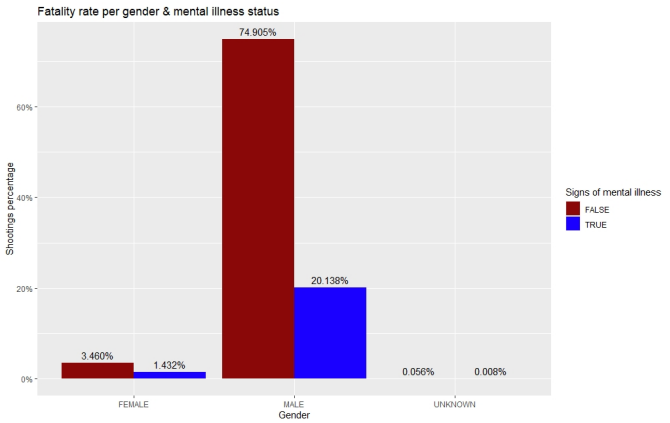

Fig. 6. The majority of fatal police shootings include males and mentally health people.

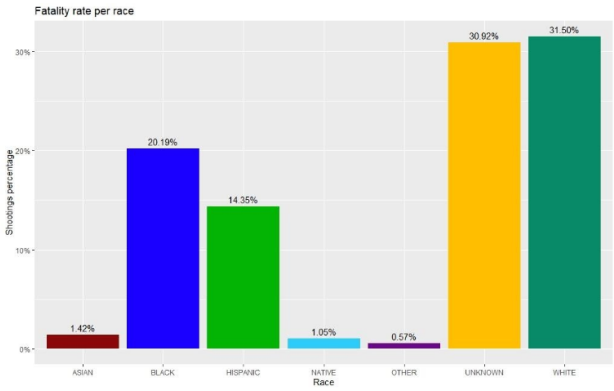

Fig. 7. With 31% being unknown, we can't conclude on which is the most targeted race of fatal police shootings.

VI. Conclusion

In this paper it was shown that:

1) the number of fatal police shootings in the US:
   - in 2017 it was 4x higher than in 2000.
   - was higher for males.
   - was higher for people with no sign of mental illness.
   - was higher for adults.
   - was higher for white people than other race, excluding the fact that a big percentage was not reported.
   - included a plethora of distinct weapons with the 3 most frequent being gun, knife and unarmed.
2) many entries failed to report used weapon, race and age.
3) many cities did not provide any statistics.
4) there were many cities exclusive to one race.
5) fatal police shootings involved 1-year-old children to 107-year-old elders.

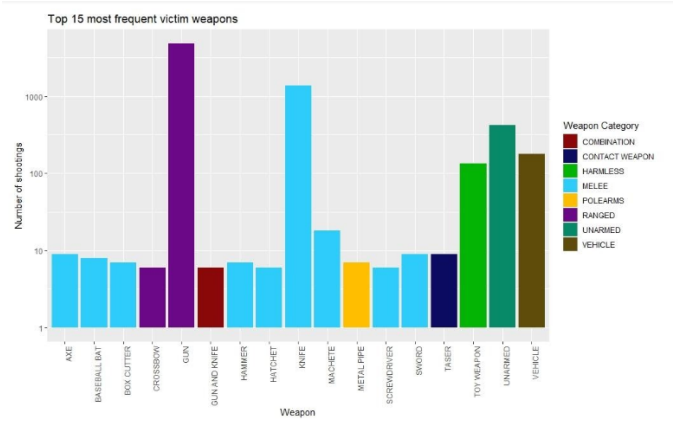

Fig. 8. These are the 15 most used known weapons carried by victims of fatal police shootings.

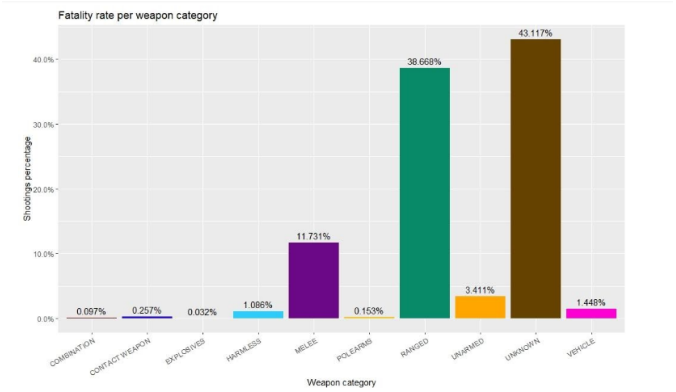

Fig. 9. The majority of entries in the dataset fail to report whether a weapon was carried or not by fatal police shootings' victims.

Further data exploration by combining dimensions can reveal more conclusions, on which specific type(s) of people were targeted in this 18-year period.

References

[1] *CRISP DM*. Wikipedia. https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining. – Accessed on 05.07.2020

[2] Awram, Chris: *US Police Involved Fatalities*. https://data.world/awram/us-police-involved-fatalities. – Accessed on 05.07.2020

[3] Karolina Wullum: *Fatal Police Shootings in the US*. https://www.kaggle.com/kwullum/fatal-police-shootings-in-the-us. – Accessed on 05.07.2020