Police Fatal Shootings in USA

Abstract - Our project is based on CRISP-DM Model. After finding the proper data sets for our ptoject we went to Business Understanding phase, where we defined the goals and criterias 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 hormonized, we have created CDWH with dimensions and fact tables. For analytics purposes we made a Data Mart out of our CDWH and perfromed the analysis. The key findings of the analysis are the amount of killings based on race, weapon, time, location and age. The goal of the project has been reached and and the result can be found in the analytics part.

Introduction -

1. Scenario:

Our goal in this project report is to show how social conditions such 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. Our main resources are 7 csv files that contain the data to be analysed. This data will be harmonized through the Staging Area before is inserted into the Database. The main tools used during the process were PHPAdmin, MYSQL Workbench, XAMPP, Notepad++ and RStudio.

1. 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 datas and some of graphics that represent these relationships visually. Finally, we conclude our work in SectionVI with the result of our analysis and what meaningful information we could get from it.

II. Open Data: MyDataSourceX

III. Data Preparation

1 - Weapon mapping:

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 “US\_Police\_Fatalities\_Stage\_Map”.

IV - Concept for the CDWH

1 - CDWH’s weapon dimension:

Weapon dimension is 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”.

2 - CDWH’s time dimension:

The script used for the time dimension tables is a simplified version of the scrip written by Prof. Dr. Markus Goldstein. A procedure is used to build the calendar for days, months and year.

3 - CDWH’s mentalillnes table:

A True / False value is used to determine the mental state of the victim.

4 - CDWH’s race table:

Describes to which race the victim belonged to. The values are: white, black, hispanic, native, asian, other or unknown.

5 - CDWH’s gender table:

Describes the gender of the victim: male, female or unknown.

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

6 - CDWH’s Fatalities table:

The fatalities table is a fact table in our CDWH. Fact table is a primary table in a dimensional model, which contains foreign keys to another dimensions and a fact. The fact of our Data Warehouse is number of shootings.

The 4 remaining dimensions were aggregated as follows:

1 - DM’s time dimension: TYear, TMonth and TDay were aggregated into TDay.

2 - DM’s age dimension: AgeGroup and Age were aggregated into Age.

3 - DM’s weapon dimension:Weapon and WeaponClass were aggregated into Weapon.

4 - DM’s location dimension: City and State were aggregated into City.

For the creation of the analysis View table, we aggregated all the information from the tables together (leaving out only the IDs). Since the Fatalities fact table was composed mostly of only foreign keys, it was used to cross the information between the different dimensions and pull out the real content that each ID represented.

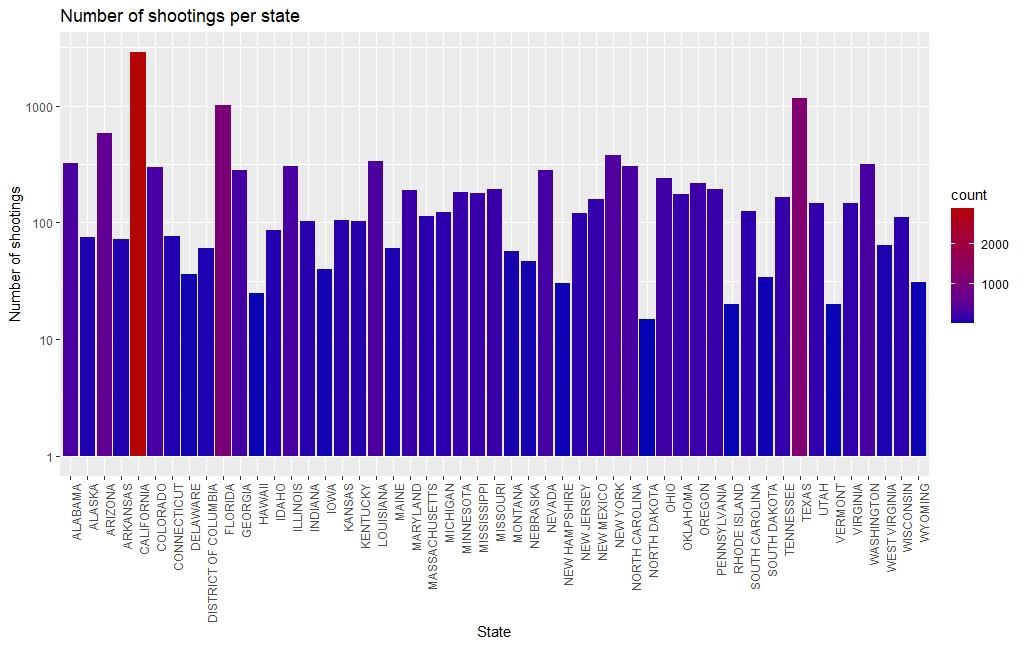
V. Analytics

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.

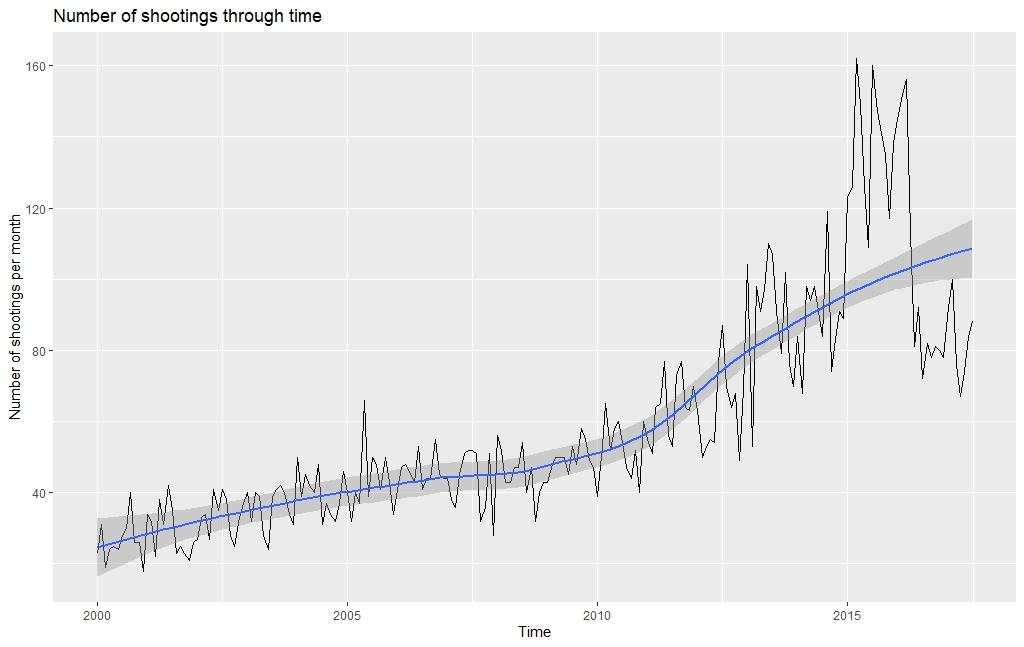
Basic summary statistics:

* Dataset consists of 12.306 entries and 21 different columns.
* Year: min = 2000, max = 2017
* CHighSchoolLevel: min = 0%, max = 100%, average = 76,52%
* CPovertyLevel: min = 0%, max = 76,40%, average = 18,03%
* CHouseholdIncome: min = 0$, max = 194.336$, average = 45.490$
* CShareWhite: min = 0%, max = 100%, average = 58,39%
* CShareBlack: min = 0%, max = 98,6%, average = 15,03%
* CShareNativeAmerican: min = 0%, max = 99,7%, average = 1,27%
* CShareAsian: min = 0%, max = 67,1%, average = 4,8%
* CShareHispanic: min = 0%, max = 98,2%, average = 22,08%
* Age: min = 0, max = 107, average = 34

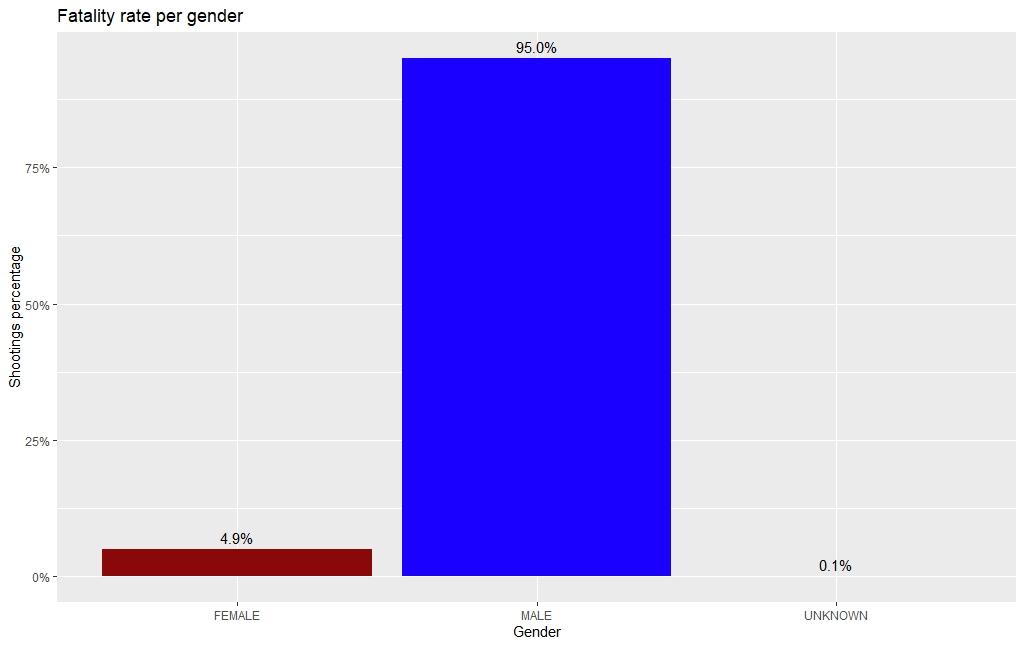
Data visualization:



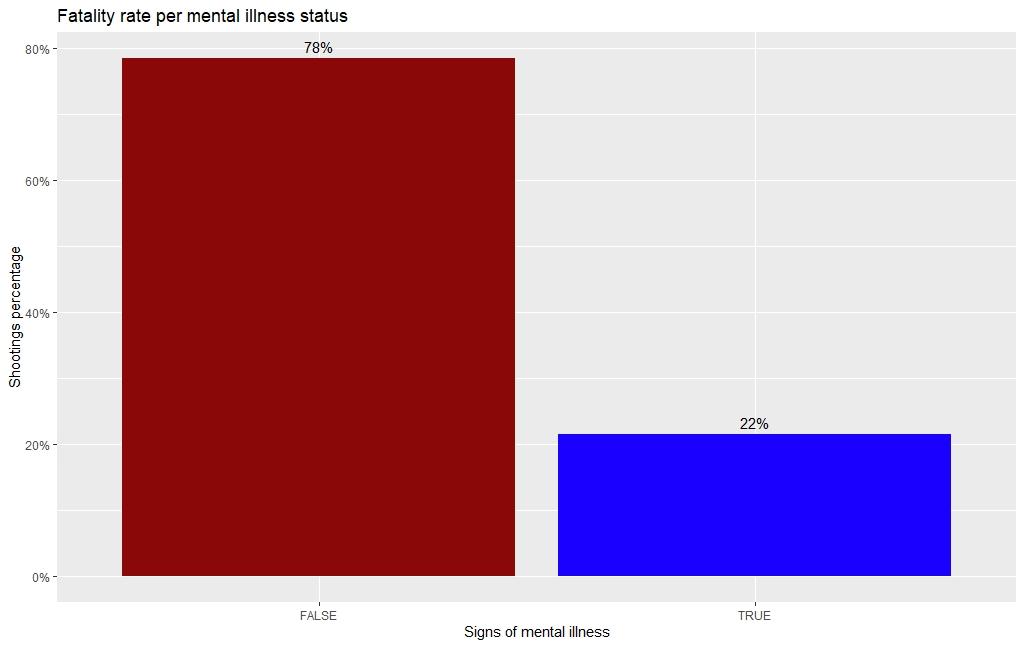
It is clearly shown that the states with the highest number of police fatality shootings are California, Texas, Florida, etc.



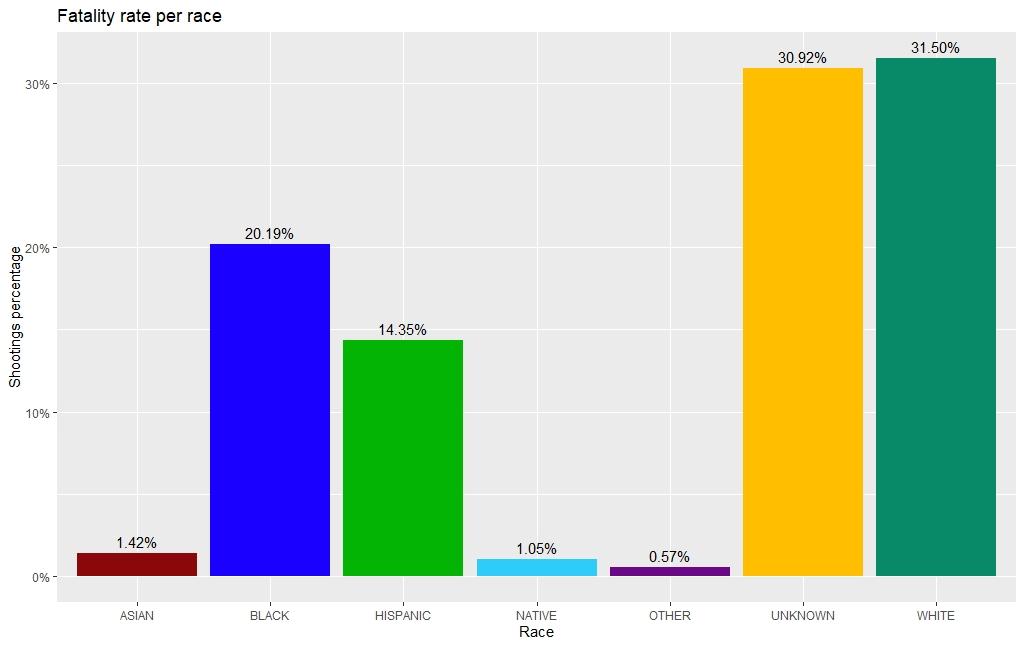
Monthly reported fatal police shootings have increased through time, with the biggest spike reported shortly after the 2016 presidential election.



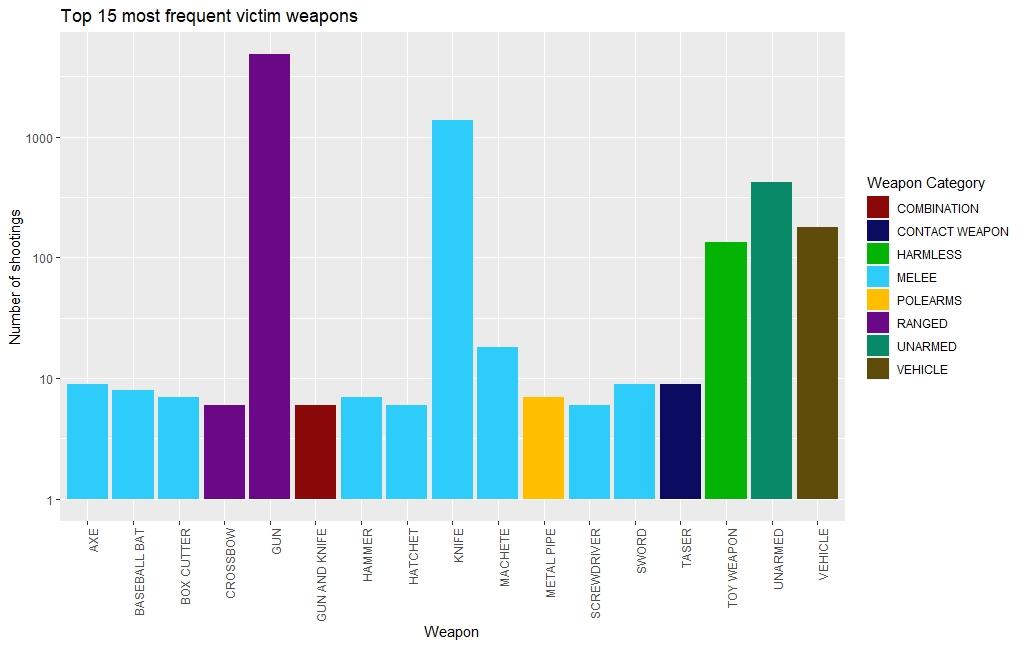
The majority of fatal police shootings include males.



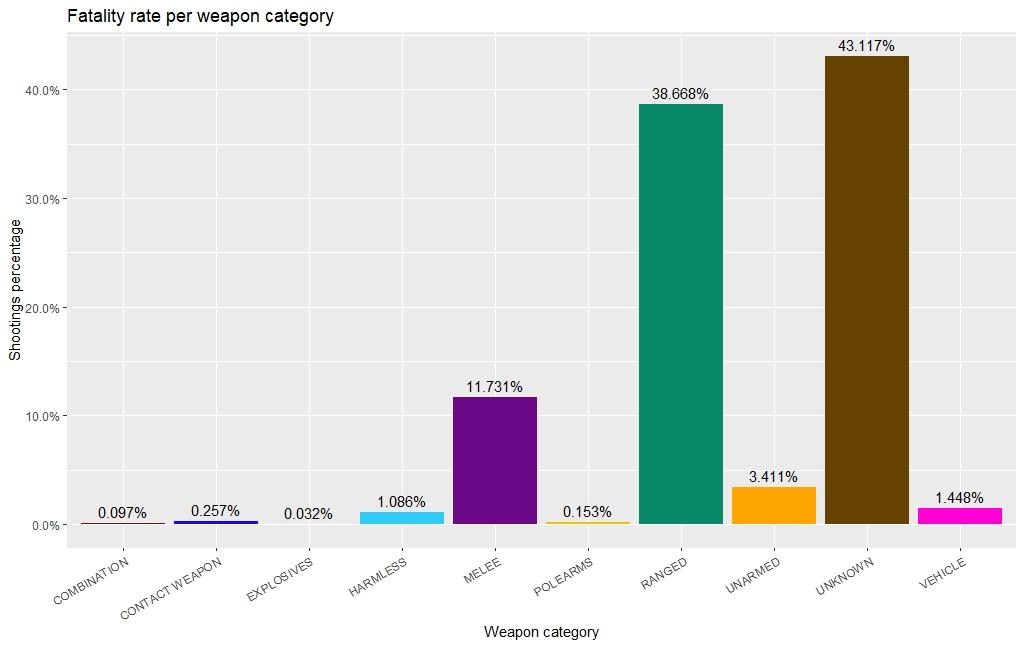
Most of the victims of fatal police shootings had no sign of mental illness.



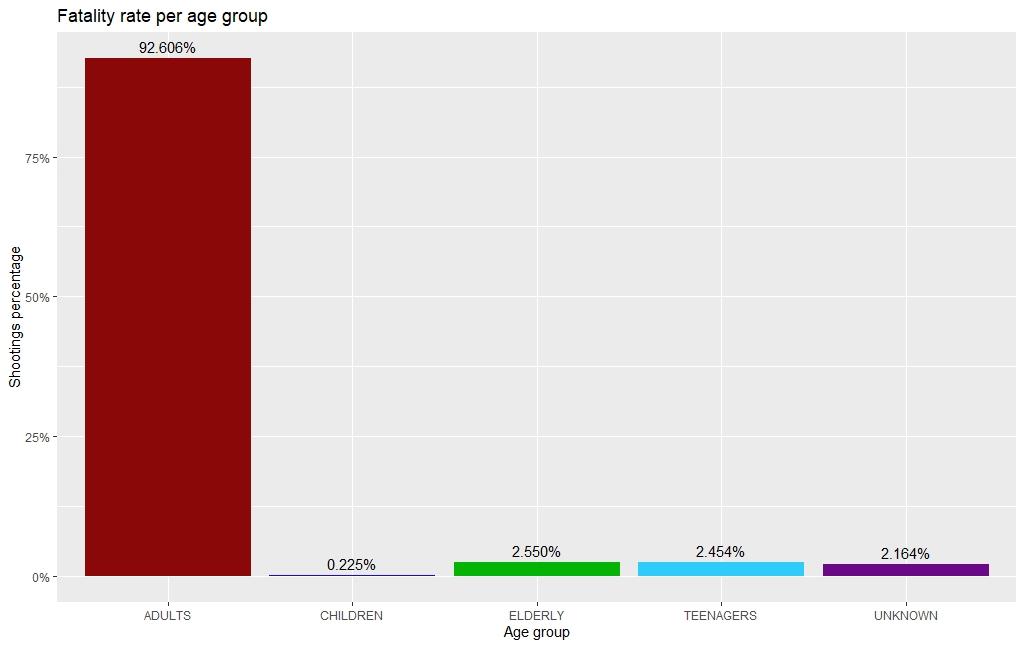
With ~31% being unknown, we can’t conclude on which is the most targeted race of fatal police shootings.



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



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



The majority of targeted age group involved in fatal police shootings is adults.

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.

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