Enri Täär, Patrik Pruunsild

# A10 Report

https://github.com/Enri-Taar/PoliceDS

## Business understanding

### The background, goals and success criteria

As US police shootings have been a topic of conversation for quite some time, The Washington Post decided to log every fatal shooting by an on-duty police officer in the United States from year 2015 to year 2020. The data that was gathered in this 5 year span will be used to carry out not a conclusive, but rather a descriptive analysis of the fatal shootings by the US police.

This data analysis project will be carried out for both an organization dealing with the general welfare of citizens and the United States' Federal Law Enforcement's Department of Homeland Security. The overall aim of this project is to patternize the characteristics of police shootings that have taken place between 2015 and 2020 in the United States and give some insight to both aforementioned organizations whether there may exist general issues regarding these shootings that need further investigation and potentially legislative, jurisdictive changes that should require imposing by the higher hierarchies of the federal government of the United States. The results of the project at hand could also possibly impact the general take of the organization dealing with the welfare of US citizens by mapping pre-emptive measurements to protect people bearing the characteristics most dominantly exhibited in the shootings data.

The data analysis project will be deemed successful if descriptive information can be extracted, which can lead both of the aforementioned organizations to the possession of a better understanding of the potential problematic scenarios in these shootings. However it must be said that no conclusive arguments can be extrapolated solely on one dataset that only describes the shootings that have taken place, not all the other police and citizen interactions that have not come down to fatal shootings by the police. However, taking this into fact, mapping out potentially dominant scenarios can lead the organizations to the conduction of more comprehensive research, which then could possibly give conclusions on any predispositions.

**Resources**

The requirement for this project is that it should be finished before the 17th of December.

There is always a risk of potential delays in the project but the project result carried out by the group with it's disciplinary approach will not be affected. If it happens that one of the two members in the group has any issues regarding making progress with the project, a contingency procedure of temporary higher workload for the other one will be imposed.

All terminology that needs further explaining will be delivered throughout the realization of the project. The aim of our group is not to rely on complex and/or ambiguous terms to avoid any misunderstanding in the interpretation of our results.

There will be no monetary or material benefits or costs regarding this data analysis project. The only benefit for the group would be good evaluation points.

**Defining data-mining goals and success criteria**

The goal of the data-mining process is to deliver a visually pleasing poster for presentation purposes which is easy to read, yet detailed enough. Any models that we intend to build during the mining process will also be considered as deliverables. Code for data analysis will also be delivered in the group's github repository.

Data-mining will be deemed successful if all models and reports generated through the process will give descriptive results as a whole which support the aims set under business success criteria. We hope to find strong correlations between these datasets and build predictive (regression) models with statistical significance which would point out any *potential* causality.

## Data understanding

### Gathering data

For our goals, the data has to be of different kinds, booleans, dates, strings, numeric. All these will account to something whilst data-mining. The required data does exist and it is all highly usable. The data used by our team is gathered by the Washington Post, who have made it available on their github. The source is highly trustworthy and quite good, as they update their database on a regular basis. We found a slightly edited version of their dataset on Kaggle, in which 3 columns have been dropped, which play a small role in our research, those being latitude, longitude and is_geocoding_exact. These 3 variables would play no role for our team's research. We will also drop the name column from our research, as this would play no role. We have tested the data to check if it can be worked on.

### Describing data

For data analysis our group has a dataset of nearly 5000 fatal shootings by the US police that have taken place between 2015 and 2020. Each row in the dataset describes the following information: name of shot person(string), date of shooting(date), manner of death (shot or shot and tasered)(string), whether the person was armed(string), age of shot person(numeric), gender of shot person(strings), race of shot person(strings), location (city) of event(strings), location (state) of event(string), signs of mental illness for shot person(boolean), threat level (whether the person shot attacked the officer)(string), whether the person fleed(string), whether the officer had a body camera(boolean) and with what (if was) was the person shot armed with (string) (almost 4895 rows, 14 columns).

There are some constraints as to interpreting the validity of some statistics. For example threat level (attacking) can be quite a subjective assumption, as one officer may describe the person running towards them as a person with an intent to attack and another officer may not. The exact same subjectiveness goes along with the assessment of possible signs of mental illness, as the person shot could have been under the influence of some sort of substance or the officer could have made a wrong assumption in general. However, as this statistic is described with the word “signs”, it implies that it’s not factual but subjective. The potential subjectiveness of some parts of the data will be taken into account in the description of results, as the subjectiveness is also a risk with detrimental potential in finding valid correlations.

## Exploring data

Quality problems - there are 51 states represented in the dataset whilst the US has 50 states, Washington DC is counted twice as WA and DC. We also noted that some entries have been removed from the dataset, because the ids skip over some values.

For data preparation, we have to fix the quality problems, the states issue can be easily resolved. Whilst the id issue, if we keep the ids as they are, it would be easier to update the results, when getting a newer dataset from the official Washington Post github, as their dataset is updated quite often.

## Verifying data quality

There actually aren't many issues regarding data quality, all the issues that we have found so far can be fixed by either changing the data or dropping the variable in question entirely. Even if we have to drop some data, we will still have enough to conduct our research.

## Project planning

1. Project setup and initial exploration of data. As we have only had the opportunity to explore the data by making sure it's quality standards are met, we will initially have to do further exploring of the data to get the project started. We will possibly be running different sorts of analyses and doing further research on things that catch our eye. This will most probably take up to 15 hours, 7.5h from both group members. **(Total 15h)**
2. We will discuss what has caught our eye and will plan in which direction we will head with the actual data-mining. This will most probably take up to 3 hours. **(Total 18h)**
3. Actual data-mining, which includes the construction of various models, running some interesting analyses etc., writing up results and documenting. This will most probably take up to 30 hours, 15h from both group members. **(Total 48h)**
4. Collection of work done by both group members, discussing and reviewing what both of the members have done and what should need modifications or perhaps removal. Cleaning up the code. This will probably take up to 7 hours. **(Total 55h)**
5. Designing of the poster, discussion of what and how should be highlighted there. **(Total 60h)**