



UNIVERSITEIT VAN AMSTERDAM

DATA ANALYSIS AND VISUALIZATION

KUNSTMATIGE INTELLIGENTIE

---

# Visualisatie en Analyse vuurwapengeweld in Amerika

---

Willem van der Spek

Joris Hijstek

Bram Bakker

Silvan Murre

June 28, 2018

## Inhoudsopgave

<b>1</b>	<b>Inleiding</b>	<b>2</b>
1.1	Motivatie . . . . .	2
1.2	Onderzoeksvragen, hypothese en verwachtingen . . . . .	2
<b>2</b>	<b>Methode</b>	<b>3</b>
2.1	Data pre-processing . . . . .	3
2.2	Exploratory Data Analysis . . . . .	3
2.3	In-depth analysis . . . . .	5
<b>3</b>	<b>Resultaten</b>	<b>5</b>
<b>4</b>	<b>Discussie</b>	<b>9</b>
4.1	Conclusies . . . . .	9
4.2	Evaluatie . . . . .	10
<b>5</b>	<b>Bijlagen</b>	<b>11</b>

# 1 Inleiding

## 1.1 Motivatie

14 Februari 2018. De datum waarop Nikolas Cruz in Parkland, Florida het vuur opende op zijn medeleerlingen en leraren. 17 mensen kwamen om, en 17 anderen raakten ernstig gewond. Het werd de dodelijkste 'school shooting' ooit in de Verenigde Staten. Als gevolg hierop barstte in Amerika de discussie over wapenwetgeving opnieuw los, en werd het nationaal bespreekpunt nummer een. Scholieren en leraren organiseerden massale 'marches', zoals de 'March for Our Lives' en op de meeste nieuwskanalen waren experts met een mening te zien. Het nationale Amerikaanse debat heeft onderzoek over wapengebruik hevig gepolitiseerd. Sommigen vinden dat het hebben van een wapen een recht is, anderen vinden dat het niet veilig is als elke burger een wapen mag dragen. Daarom is het interessant om een stap terug te nemen en objectief de ruwe, objectieve data te bekijken, zoals een volledige dataset van elk schietincident in de Verenigde Staten van de afgelopen vier jaar, en deze dataset te analyseren.

## 1.2 Onderzoeksvragen, hypothese en verwachtingen

De dataset wordt geanalyseerd met behulp van verschillende onderzoeksvragen. Er wordt gekeken in hoeverre de striktheid van de wapenwetgeving in verschillende staten invloed heeft op het aantal schietincidenten per capita in die staat. Ook wordt gevraagd op welke manier de hoeveelheid schietincidenten verschilt over de tijd, bijvoorbeeld op welke dagen wordt het meest geschoten en in welk seizoen? Daarnaast is een van de onderzoeksvragen een demografische, met het oog op de hoeveelheid 'school shootings' in Amerika is het interessant om te onderzoeken welk aandeel van massale schietpartijen door jongeren wordt uitgevoerd en of dit significant hoger is dan het aandeel wat jongeren hebben in normale schietincidenten. Om een duidelijker beeld van wapengebruik in Amerika te scheppen wordt verder nog de relatie tussen slachtoffer en dader in schietincidenten onderzocht.

De hypothese op de onderzoeksvraag van de wapenwetgeving per staat is simpel: hoe strikter de wapenwetgeving is, hoe minder doden en gewonden per capita. Er zullen dagen van de maand zijn waarop meer incidenten plaatsvinden dan anderen, ofschoon er nog geen uitspraak wordt gedaan over specifieke dagen. Bovendien wordt er meer geschoten in de zomermaanden en minder in de wintermaanden, over de tijd zal het aantal schietincidenten dalen aangezien de wapenwetgeving in

Amerika over het algemeen striktere wapenwetgeving aan het invoeren is. Massale schietpartijen waarbij 4 of meer mensen om het leven komen worden relatief door jongeren uitgevoerd, omdat ze emotioneel instabieler zijn dan ouderen. Voor het onderzoek tussen relaties zal blijken dat wapenincidenten vaker gerelateerd zijn aan bendes aangezien bendes meer toegang hebben tot wapens. Naast al deze vragen en hypotheses is het altijd belangrijk in dit soort onderzoek naar datasets om een oog open te houden op opvallende datapunten en die wellicht verder te onderzoeken.

## 2 Methode

### 2.1 Data pre-processing

In de data pre-processing fase was de data gecleaned voor verder gebruik. Dit was relatief simpel aangezien de gekozen dataset al in csv-formaat gegeven werd. De manier waarop de informatie aan elkaar gekoppeld werd moest omgezet worden naar data waar Python iets mee kon. Om dit te bereiken werden de columns helemaal gelezen en door seperatoren weg te halen kon alle data in dictionaries worden gezet. Ook werd in deze fase elke kolom die niet nodig of nuttig was voor het onderzoek verwijderd.

### 2.2 Exploratory Data Analysis

Een van de vragen die gesteld werd over de dataset ging over opvallende verschillen in wapengeweld tussen staten. Uit de hypothese kwam naar voren dat als een staat een striktere wapenwetgeving hanteert, er minder doden vallen. In de dataset stond niks over de striktheid van de wapenwetgeving. Hiervoor moest dus een 'striktheids-score' ontwikkeld worden gebaseerd op de hoeveelheid regulaties die de desbetreffende staat heeft. De schaal ging van 1 tot 6, omdat er naar 6 verschillende wapenwetten werd gekeken. Deze 6 wapenwetten zijn de belangrijkste wetten, omdat voor elke staat de mate waarin de wapenwet wordt toegepast geregistreerd staat. De wetten luiden als volgt:

1. Achtergrondcontrole bij aankoop wapen.
2. Preventie van toegang tot wapens bij kinderen.
3. Toestemming van het dragen van verborgen wapens in het openbaar.
4. Het verbieden van verkoop van wapens aan daders van huiselijk geweld.

5. Extreme Risk Protection Order (ERPO): Families en wethandhaving kunnen een verzoek aan de rechtbank doen om wapens weg te halen bij mensen die zichzelf of andere mensen pijn kunnen doen.
6. Verbod op militaire wapens.

Per wet kon voor elke staat de volgende score worden toegekend:

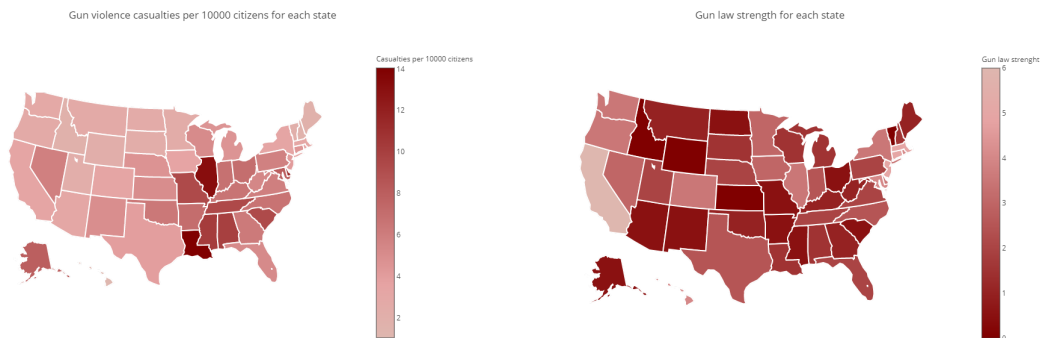
1. Geen wet/Geen handhaving van de wet - 0 punten
2. Partiële wet/Matige handhaving van de wet - 0,5 punten
3. Strenge wet/Sterke handhaving van de wet - 1 punt

Voor alle 51 staten werd een nieuw csv-bestand gemaakt waarin alle benodigde data stond. Door het totaal aantal slachtoffers van wapenincidenten in een staat door de complete populatie in die staat te delen en te vermenigvuldigen met 10.000 werd het belangrijkste gegeven berekend. Met behulp van het csv-bestand werden twee heatmaps geplot: een heatmap bestond uit het aantal casualties per staat per 10.000 inwoners en de andere bestond uit de striktheid van de wapenwetgeving per staat. Met de twee heatmaps werd een duidelijk beeld weergegeven over de invloed van regulaties op schietincidenten. Omdat elk schietincident gelinkt is aan een datum konden verschillende lijnen geplot worden die de hoeveelheid schietincidenten verdeeld over de tijd weergeven. Hierdoor werden meer vragen opgeroepen die terugkomen in de in-depth analysis. Om de vraag over het aandeel van jongeren in massale schietincidenten te beantwoorden moesten eerst wat cijfers verkregen worden. Door per incident waar een jongere in voorkomt, die jongere te matchen met de rol suspect, en hierna te kijken of het een mass shooting was konden de juiste cijfers berekend worden. Hiermee werd een stacked bar plot gemaakt. Voor de analyse tussen relaties werd elk voorkomen van een bepaalde relatie geteld en in een pie chart gezet. Ook werd een bar chart gemaakt die per staat bepaalde vormen van incidenten in kaart bracht, wat weer handig vergeleken kon worden met de twee heatmaps.

## 2.3 In-depth analysis

Gezien het gebrek aan continuïteit van de variabelen binnen de dataset is er maar één visualisatie gedaan aan de hand van *In-depth analysis*. Er is gekozen voor een grafiek van punten, waarin op de horizontale as de leeftijd van daders stond en op de verticale as het aantal doden en gewonden die door deze dader waren aangericht. Alle datapunten waar een dader bekend was zijn op deze wijze geplot, mits hun leeftijd niet groter was dan die van de oudste mens ter wereld. (Om eventuele foutieve datapunten uit te sluiten) Vervolgens is met de *SKlearn*-module van Python regressie toegepast volgens de kleinste-kwadraten methode. De parameters die zijn gebruikt voor de visualisatie zijn de richtingscoëfficiënt en de hoogte voor deze regressie. Naast de visualisatie is er statistische analyse uitgevoerd voor de eerste dag van de maand. Het gemiddelde op alle eerste dagen van de maanden exclusief januari is berekend en het gemiddeld aantal incidenten op een willekeurige dag is vervolgens als normaalverdeling beschouwd. Het gemiddelde aantal incidenten voor de eerste dag van de maand buiten januari is hierop ten slotte getoetst.

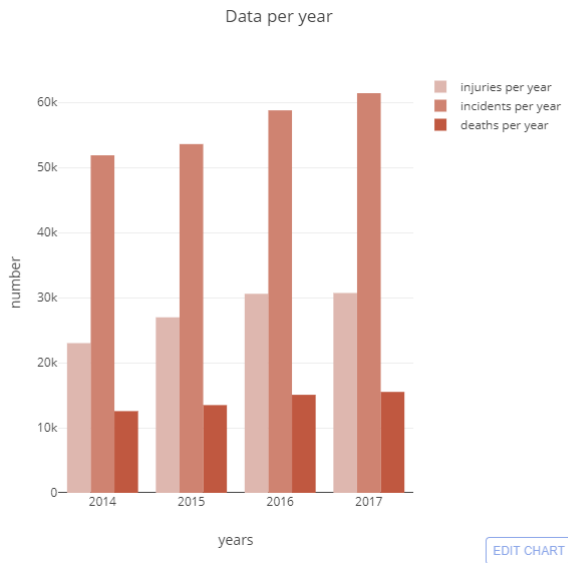
## 3 Resultaten



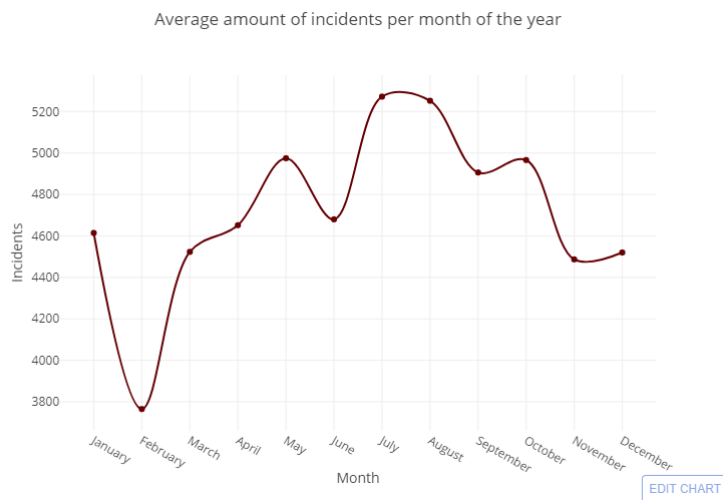
Over het algemeen lijkt wapenwetgeving te helpen als we naar de twee heatmaps kijken.

## Visualisatie en Analyse vuurwapengeweld in Amerika

---



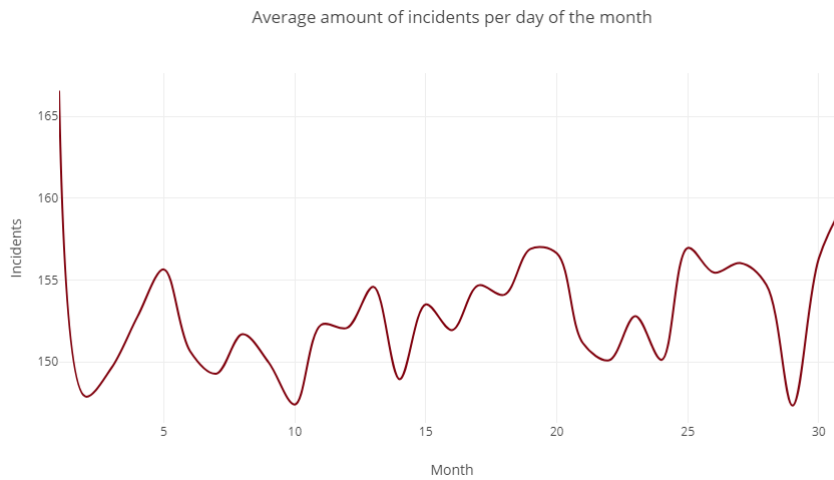
Over de afgelopen vier jaar hebben er meer incidenten plaatsgevonden met betrekking tot vuurwapengeweld.



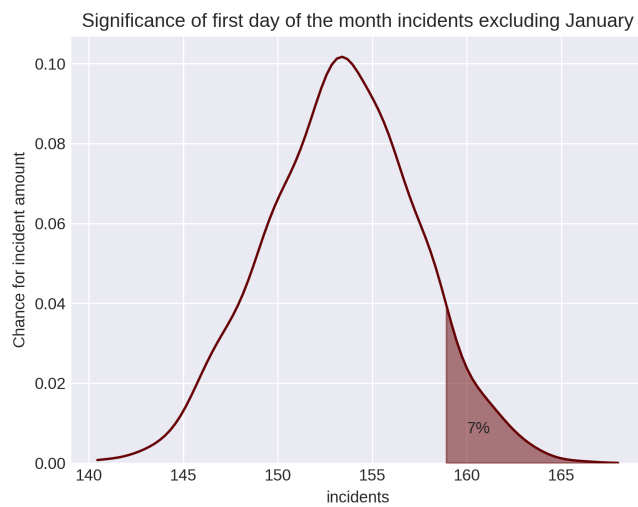
In de zomer wordt het meest geschoten en in februari het minst.

## Visualisatie en Analyse vuurwapengeweld in Amerika

---



Op de eerste en laatste dag van de maand wordt ook significant meer geschoten.

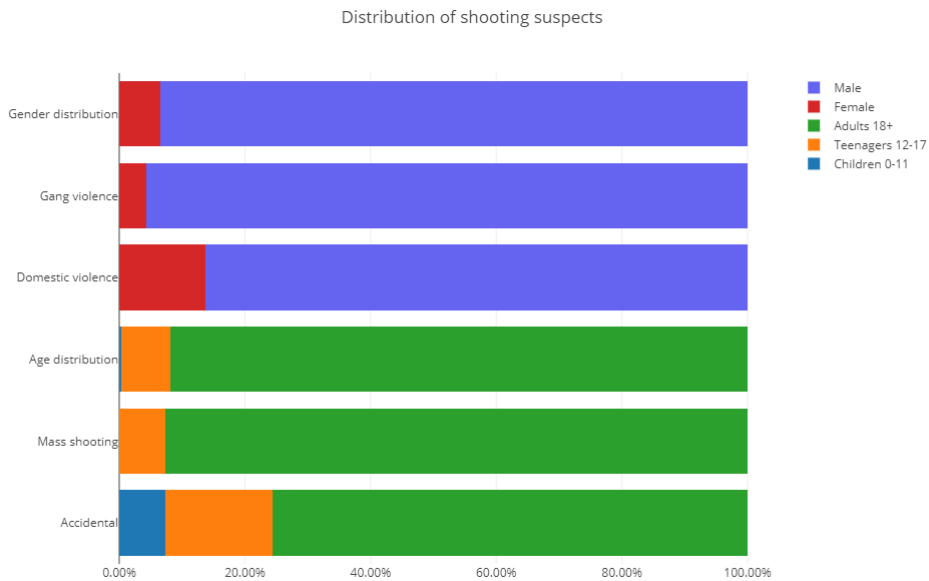


Statistische analyse duidt erop dat de eerste van januari de hoeveelheid incidenten zo hoog maakt.

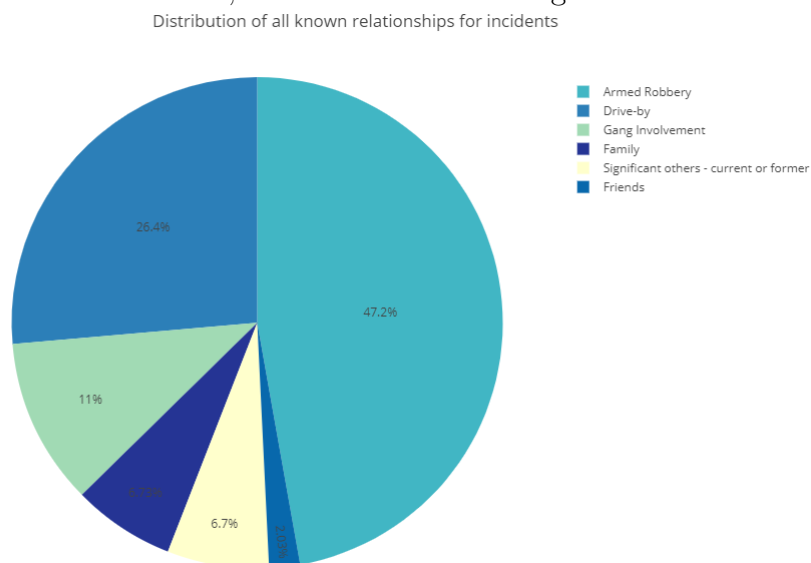


## Visualisatie en Analyse vuurwapengeweld in Amerika

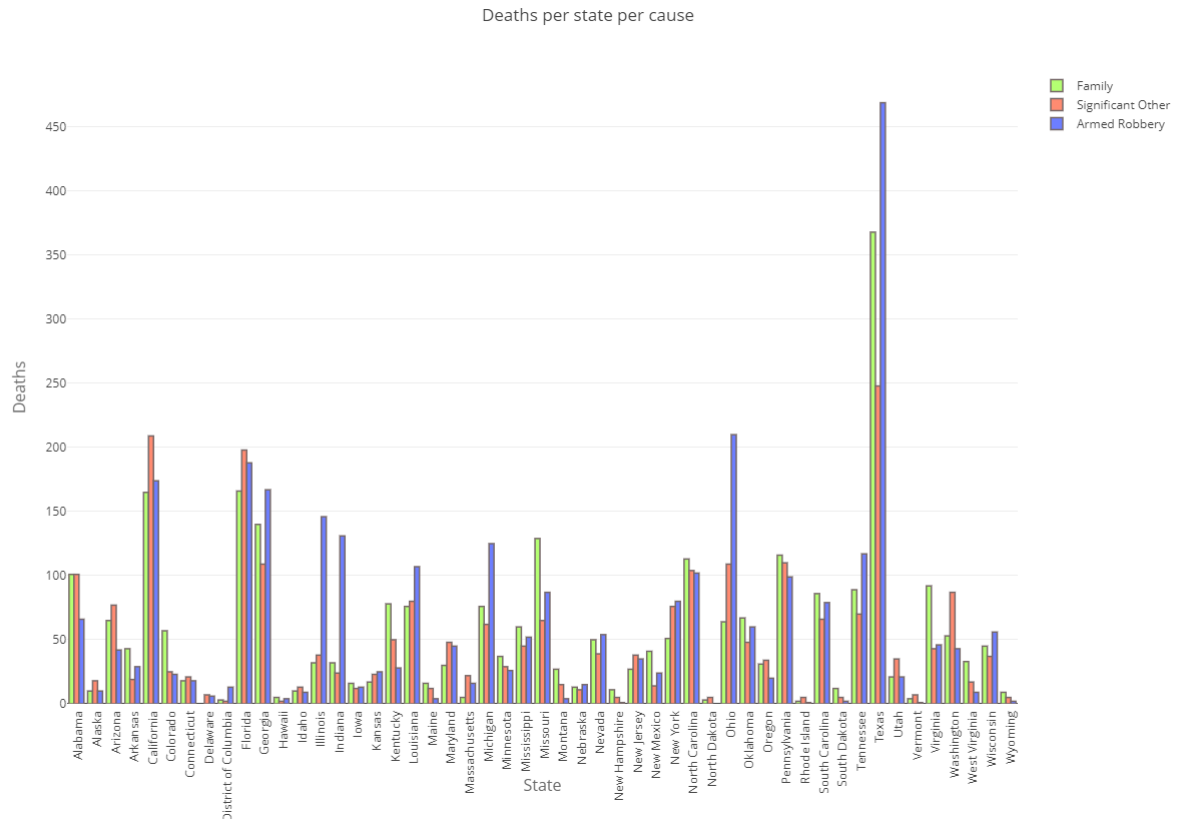
---



Jongeren plegen ongeveer hetzelfde aandeel van massa schietincidenten als in normale schietincidenten, maar behoren tot een groter deel van de schiet accidenten.



Het grootste deel van schietincidenten betreft armed robberies



In texas zijn er blijkbaar extra veel armed robberies

## 4 Discussie

### 4.1 Conclusies

De vraag of een losse wapenwetgeving voor een staat correleert met meer doden en gewonden wordt aan de hand van onze resultaten bevestigd. Onze hypothese dat een lossere wapenwetgeving minder doden en gewonden bewerkstelligt is dus correct. Verder bevestigt de data dat er in de zomer de meeste incidenten zijn en in de winter over minder vuurwapengeweld wordt gepleegd. Voor de dagen van de maand geldt dat de eerste dag van de maand significant meer incidenten kent dan de andere dagen van de maand. Ondanks het feit dat dit door de nieuwjaarsdagen van de afgelopen jaar wordt bewerkstelligt, is de hypothese voor het aantal incidenten voor de dagen van de maand formeel gezien juist, en zal deze moeten worden bevestigd. De hypothese dat er in de afgelopen aantal jaar minder vuurwapengeweld zou

plaatsvinden is aan de hand van de resultaten daarentegen totaal misplaatst en zal dus moeten worden verworpen. Daarnaast werd gehypothetiseerd dat bendes het gros van het wapengeweld zouden vertegenwoordigen. De data duidt er echter op dat andere relaties markanter zijn en hiermee de hypothese tenietdoen. De laatste onjuiste hypothese is degene die jongeren als hoofdverdachten van massale schietpartijen beschuldigde. Het blijkt dat jongeren niet significant vaker verdachten zijn van massale schietpartijen dan normale schietpartijen, verder blijkt na meer onderzoek dat ze wel een relatief groter aandeel hebben in ongelukken. Jongeren blijken dus eerder onhandig dan volwassenen.

## 4.2 Evaluatie

Na het voltooien van het onderzoek over leeftijdsdistributie kon hetzelfde gedaan worden op basis van geslacht. De data analyse met de *stacked bar charts* is hiervoor uitgebreid. Regressie is moeilijk voor onze dataset met weinig continue variabelen. *T-distributed Stochastic Neighbor Embedding* is bijvoorbeeld geen optie gezien er geen behoefte is aan het reduceren van de dimensionaliteit van de data. Voor een volgend onderzoek is het een idee om de dataset met andere datasets samen te voegen om zo tot meer mogelijkheden tot datavisualisatie te komen. De methode waarop data van aan het incident gerelateerden werd verzameld was met behulp van een functie die het datatype *string* veranderde naar het pythonische datatype *dictionary*. Dit kon gebeuren omdat de strings dusdanig waren gestructureerd dat dubbele dubbele punten voor *binders* stonden en dubbele *pipelines* voor seperatoren stonden.

Deze documentatie leek op het eerste gezicht te gelden voor de hele dataset. Gedurende het onderzoek bleek echter dat de documentatie niet klopte voor alle rijen binnen de csv, waarop moest worden geanticipeerd. Het probleem bleek dat in plaats van dubbelvoudige dubbele punten en *pipelines*, enkelvoudige waren gebruikt voor sommige cellen binnen de csv. Deze informatie is relevant ter verbetering van toekomstig onderzoek en de dataset.

## Referenties

- [1] <https://github.com/jamesqo/gun-violence-data> James Ko
- [2] "Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2017" (XLSX). United States Census Bureau. Retrieved December 21, 2017.

## 5 Bijlagen