

Data Analyse en Visualisatie van Gun Violence in de Verenigde Staten

Nils Peters, Fabian van Stijn, Djanira dos Santos Gomes, Loes Gennissen

28 juni 2018

1 Inleiding

Door data sets te analyseren en te visualiseren kan er een goed overzicht worden gemaakt van een onoverzichtelijke massa aan data. Er is gekozen om de dataset over *Gun Violence* in de Verenigde Staten te analyseren omdat dit een interessant en belangrijk onderwerp is. De VS staat bekend als een ontwikkeld land met het meeste wapengeweld. Wapenbezit is heel normaal aangezien 42% van de Amerikaanse burgers een of meerdere wapens thuis heeft (Igielnik, 2017). Door de grote hoeveelheid wapens die beschikbaar zijn, is het niet vreemd dat er zoveel geweld plaatsvindt. Het doel van deze opdracht is het laten zien waar en wanneer er het meeste geweld in de VS voorkomt. Verder wordt er ook gekeken naar de dader- en slachtofferperspectief van de wapenincidenten. In deze dataset zijn alle gemelde wapenincidenten in de VS van januari 2013 tot maart 2018 opgenomen. Over deze dataset zijn enkele vragen.

1. Zijn er merkbare verschillen tussen de staten/steden/jaren? Visualiseer de verschillen (of gelijkenissen) die je kan vinden.
2. Wat zijn de patronen die je gevonden hebt die interessant kunnen zijn? Bevat de data enkele ongewone data die je niet had verwacht?
3. Hoe ziet het patroon tussen moordenaar/slachtoffer eruit tussen verschillende incidenten door de jaren. Is het vooral familie/relatie of zien we meer gerapporteerd *gang* geweld bijvoorbeeld?

Naast deze vragen zijn er nog enkele andere vragen bedacht om te beantwoorden.

4. Hoeveel mensen raken gewond of worden vermoord tijdens een wapen incident?
5. Zijn de mensen die bij het wapengeweld betrokken zijn vooral man of vrouw?
6. Zijn er in alle incidenten meer slachtoffers of daders?

De hypothese voor vraag 1 is dat vooral in dichtbevolkte staten die langs de kusten van Amerika liggen het meeste geweld voorkomt. In stedelijke gebieden zal meer wapengeweld voorkomen dan kleinere steden. Door de jaren heen zal naar vermoeden geen groot verschil in het aantal incidenten per jaar zijn. De hypothese voor vraag 3 is dat er vooral veel incidenten zullen zijn waarvan slachtoffer en dader elkaar niet kennen. Dit kan komen doordat in de VS nog veel gewapende berovingen en straatgeweld plaatsvindt.

De hypothese voor vraag 4 is dat er niet bij elk incident een gewonde of dode valt. Bij vraag 5 is de hypothese dat er vooral mannen bij de wapenincidenten betrokken zijn. Voor vraag 6 is de hypothese dat er meer slachtoffers zijn dan daders aangezien een dader meerdere slachtoffers kan maken met wapens.

2 Methode

Voor het programmeerdeel van de data analyse en visualisatie wordt gebruik gemaakt van Python 3. Er wordt gebruik gemaakt van Git om alle code van de de verschillende groepsleden te monitoren en de verschillende plek de code samen te voegen. Er werd gebruik gemaakt van Jupyter Notebook om in te programmeren.

De eerste stap om met de dataset te kunnen werken is de dataset te *cleanen*. De data is opgeslagen in een CSV bestand, maar niet alle data in dit bestand is compleet. Aangezien er in 2013 slechts 279 incidenten waren, is er voor gekozen om 2013 te verwijderen van de dataset. ok is er gekozen om de kolommen waarbij meer dan 40% van de data mist, te verwijderen. Dit zijn de volgende kolommen: *state house district*, *state senate district*, *congressional district*, *participant age*, *notes* en *adress*. Ook miste bij de kolom *relationships* 93.6% van de data omgezet naar *unknown*. Er stonden op sommige mensen dubbele relationships. Dus verwijderd. Aangezien er echter een vraag over de relatie van dader/slachtoffer is, is er besloten om deze data erin te laten. Er zijn 7912 rijen verwijderd die de *longitude* en *latitude* misten. Er zijn 328 rijen verwijderd bij de *incident characteristics*. Bij *guns involved* 99000 rijen omgezet naar *unknown* want er stond eerst niets. Bij 35518 rijen van *participant gender* lege plekken omgezet naar *unknown* want bij niet iedereen stond een geslacht. Bij *participant status* zijn 27001 rijen omgezet naar *unknown*. Bij *type* zijn 24290 rijen omgezet naar *unknown*. Verder is er gekozen om enkele kolommen aan te passen. De kolom met *date* is gesplitst in drie verschillende kolommen: *day*, *month* en *year*. Er zijn totaal 231122 rijen overgebleven.

Vervolgens kan de echte data analyse beginnen. Voor de analyse is met behulp van Python 3 door de resterende dataset gegaan om per onderwerp de nodige data uit de dataset te halen. Er is met behulp van Bokeh interactieve data visualisatie toegepast. Voor het geslacht wat betrokken is, de leeftijds-groepen, de rol van de betrokken mensen en de staat waarvan de betrokken mensen zich in bevinden, is een bar-chart gemaakt. Dit werd gedaan door bij de y-as het aantal personen van de desbetreffende kolom te zetten en op de x-as de verschillende variabele van de kolommen. Van de hoeveelheid gewonde mensen per incident en de hoeveelheid gedode mensen per incident is een piechart via Bokeh gemaakt. Ook voor de relatie tussen mensen die bij het incident betrokken waren is een piechart via Bokeh gemaakt.

De data van de jaren is met behulp van Python zo opgedeeld dat het aantal incidenten per jaar verkregen kan worden. Door steeds wanneer een jaartal verspringt opnieuw te beginnen met het aantal incidenten te tellen. Voor de maanden is het ook zo gedaan dat elke keer wanneer de maand met een stijgt het aantal incidenten weer vanaf nul wordt geteld. De data over de jaren is in een bar-chart gevisualiseerd en de data van de maanden in een histogram.

De plaatsen waar de incidenten hebben plaatsgevonden, zijn op verschillende manieren gevisualiseerd. Er is gebruik gemaakt van de Google Maps API waarbij de *longitude* en *latitude* werden genomen om pins op de plekken van een wapen incident te plaatsen. De database file is hiervoor van een CSV file

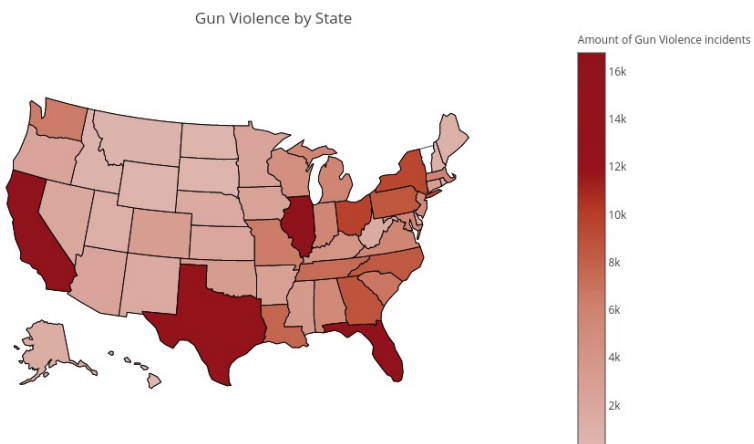
omgezet naar een JSON file want de API werkt alleen met een JSON file. Voor de steden is een bar-chart gemaakt waarbij het aantal van alle incidenten in de aparte steden werden weergegeven. Er is hiervoor gekozen om alle aparte gebieden van Chicago bij Chicago op te tellen omdat de aparte gebieden op zichzelf al een grote hoeveelheid incidenten hebben. Op de y-as staat het aantal wapen incidenten en op de x-as de steden in volgorde op meeste incidenten. Verder is het wapen geweld per staat ook gevisualiseerd. Dit is met behulp van Plotly gedaan. Er is hier een kleur schaal toegepast waardoor het verschil in aantal incidenten per staat makkelijk te zien is. Er is per jaar een kaart gemaakt met het aantal incidenten. Daarnaast is er ook een kaart gemaakt met het aantal incidenten gedeeld door het aantal inwoners van de staat. Bij de staat Washington District of Columbia kwam daar zo'n groot getal uit dat het niet mogelijk was om de andere staten daardoor goed te visualiseren. Daarom is gekozen om voor Washington District of Columbia NaN als waarde te geven.

Als laatste is een website gemaakt om alle data op te laten zien. De website bestaat uit een homepage, incidenten pagina, betrokkenen pagina en een plaatsen pagina. De website is gemaakt met behulp van HTML, CSS en Java om de grafieken van Bokeh en Plotly te kunnen visualiseren.

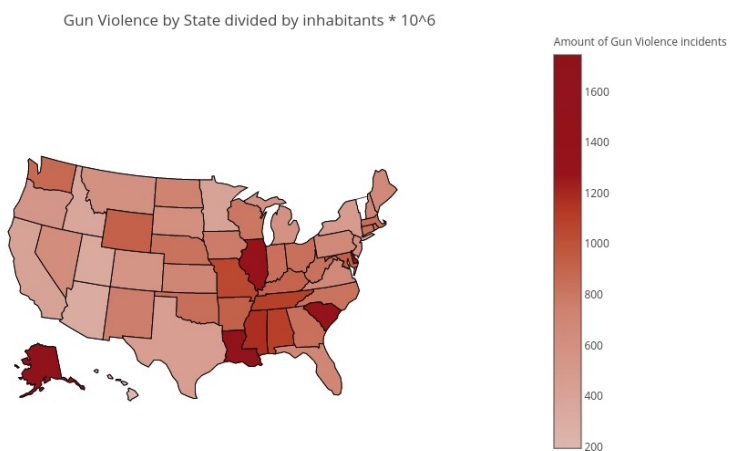
3 Resultaten

1. Verschillen tussen staten, steden en jaren

Onderstaande kaarten (fig. 1, 2) van de Verenigde Staten geven voor de jaren 2014-2018 de hoeveelheid wapengeweld per staat weer: hoe donkerder een staat van kleur is, des te meer incidenten van wapengeweld is er in deze staat gerapporteerd. Figuur 1 toont de absolute waarden, figuur 2 toont de door bevolkingsdichtheid genormaliseerde waarden.



Figuur 1.

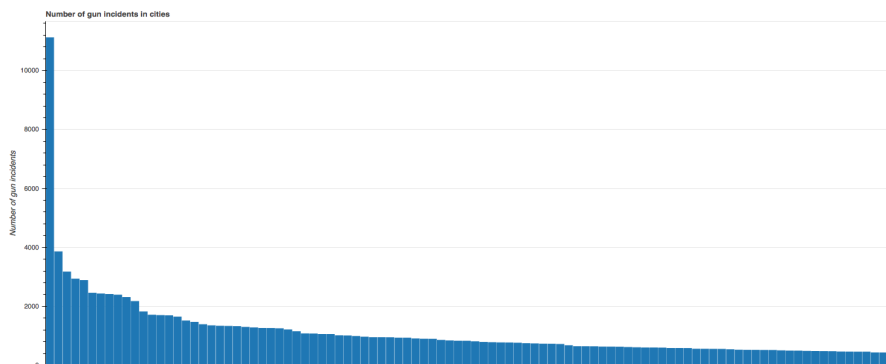


Figuur 2.

De staten met de meeste incidenten van wapengeweld blijken Texas, Florida, California en Illinois te zijn (fig. 1). Uit figuur 2 blijkt echter dat Illinois een geval apart is, vanwege het kleine aantal inwoners. Hier is dus sprake van relatief veel incidenten. Door het vergelijken van het absolute aantal incidenten

met de dichtbevolktheid van elke staat wordt bovendien zichtbaar dat ook in Alaska, Louisiana en South Carolina buitenproportioneel veel incidenten zijn gerapporteerd.

In onderstaande bar chart (fig. 3) zijn de steden in de VS te zien met het bijbehorende aantal incidenten per stad. De tien steden met de meeste gerapporteerde incidenten van wapengeweld zijn weergegeven de tabel daaronder (fig. 4). Dit zijn van links naar rechts ook de eerste tien steden op tabel 1.



Figuur 3.

Table 1: Aantal incidenten per stad

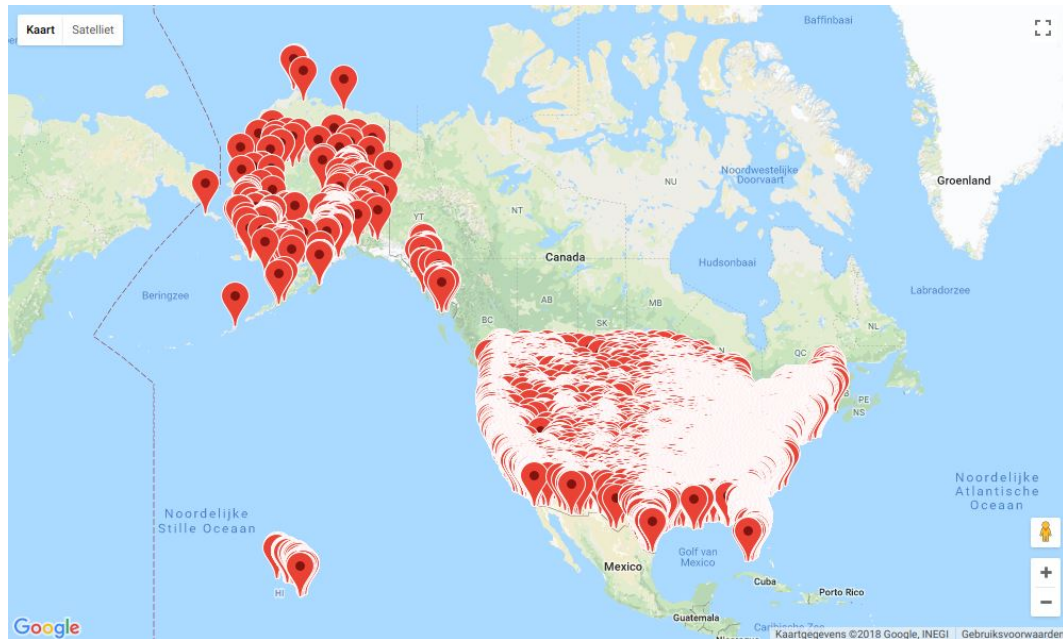
Stad	Aantal incidenten
Chicago	10278
Baltimore	3337
New Orleans	2892
Saint Louis	2427
Milwaukee	2407
Jacksonville	2369
Memphis	2300
Columbus	2134
Houston	2000
Indianapolis	1798

Opvallend is het feit dat in Chicago de afgelopen jaren significant meer incidenten hebben plaatsgevonden dan in de andere steden in de Verenigde Staten: er zijn er drie keer zoveel gerapporteerd als in de stad op de tweede plaats, Baltimore. De andere steden in de tabel hebben daarentegen vergelijkbare aantallen met Baltimore.

Verder bevinden alle tien de steden met het meeste wapengeweld zich in staten waar in totaal ook meer incidenten gerapporteerd zijn, ofwel absoluut, ofwel relatief.

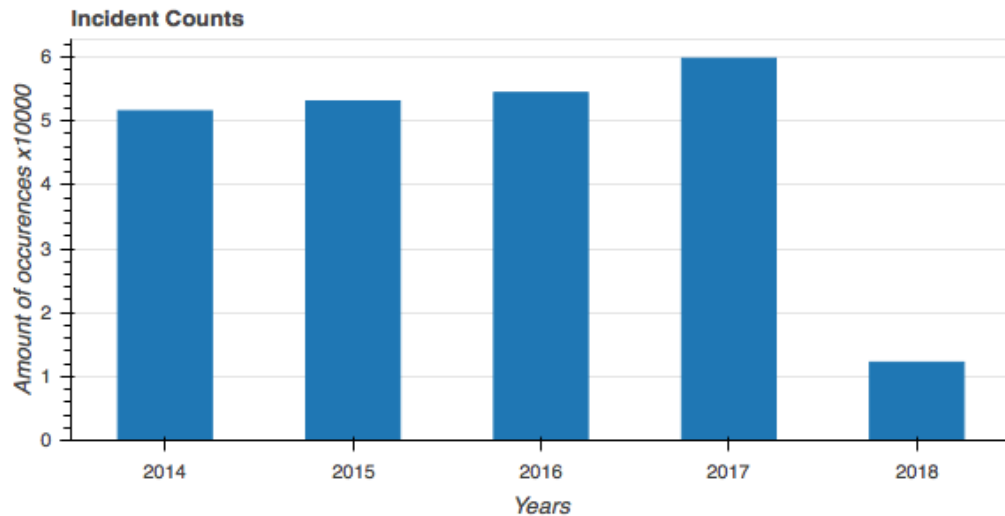
In figuur 4 zijn alle incidenten van de periode 2014-2018 op een kaart van Amerika gepinpoint; door de grote dichtheid aan incidenten op het continent

is dit echter alleen overzichtelijk voor de staten waar relatief weinig incidenten hebben plaatsgevonden.

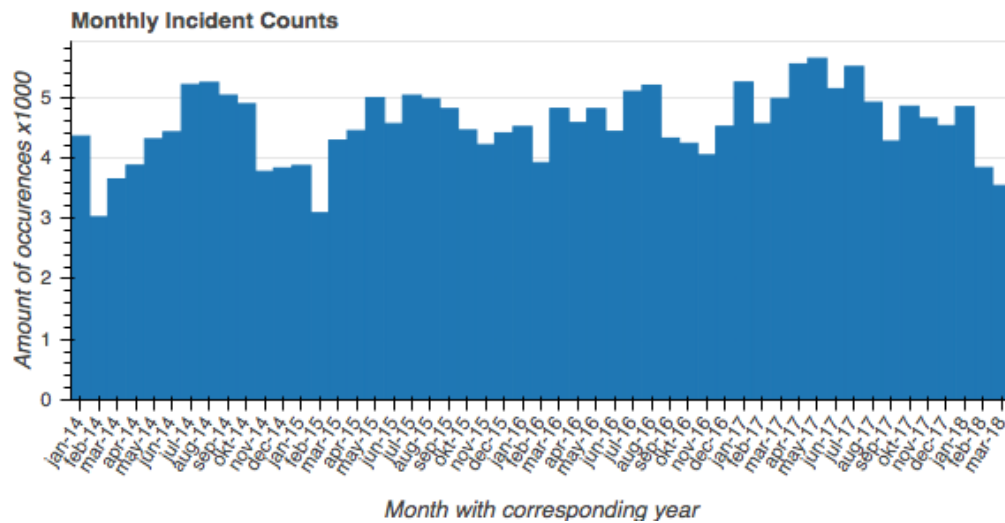


Figuur 4.

Uit figuur 4 blijkt ook hoeveel incidenten er in Alaska zijn, ondanks het feit dat het een dunbevolkte staat is. Dit wordt ondersteund door figuur 2. Onderstaande histogrammen (fig. 5,6) tonen respectievelijk het totale jaarlijkse en maandelijkse aantal incidenten in de VS in de jaren 2014-2018.



Figuur 5.



Figuur 6.

Uit figuur 5 blijkt dat het wapengeweld de afgelopen jaren gestaag is toegenomen. De hoeveelheid incidenten in 2018, waarvan alleen data zijn voor de maanden januari tot en met maart, indiceert dat deze toename ook voor heel 2018 zal gelden: in deze drie maanden waren meer incidenten dan in diezelfde periode in 2014 tot en met 2017.

Bovendien vindt er elk jaar een piek plaats in de zomermaanden (fig. 6): in juli en augustus en de maanden daaromheen vinden meer incidenten plaats dan in de rest van het jaar. Ook valt het op dat deze piek in 2017 vroeger in het

jaar plaatsvond dan de jaren ervoor, namelijk al in de lente.

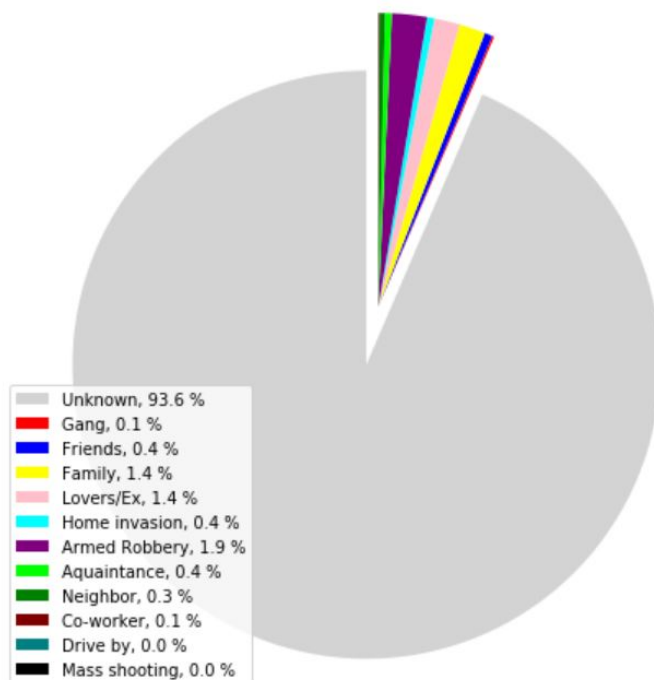
2. Onverwachte patronen en bevindingen

Zoals eerder genoemd zijn er in Chicago een onverwacht groot aantal incidenten gerapporteerd vergeleken met de andere steden (fig. 3). Dit komt terug op figuur 1: Illinois, de staat waar Chicago ligt, is één van de staten met het meeste wapengeweld per inwoner. Wat ook verrassend is, is het patroon van pieken in wapengeweld in de zomermaanden (fig. 6). Daarnaast blijkt er in absoluut opzicht meer wapengeweld plaats te vinden in steden aan de kust dan in het binnenland (fig. 1).

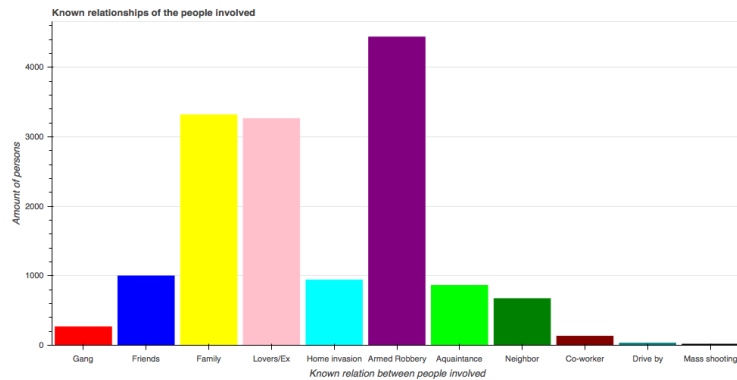
3. Killer/victim-relaties

Over de relaties tussen betrokkenen is over het algemeen weinig bekend; daarom is ervoor gekozen om te focussen op de incidenten waarvan de relaties wel genoemd zijn. Onderstaande pie chart (fig. 7) geeft weer welk deel van de totale verzameling van incidenten gebruikt is voor de bar chart daaronder (fig. 8).

Relationships between people involved



Figuur 7.



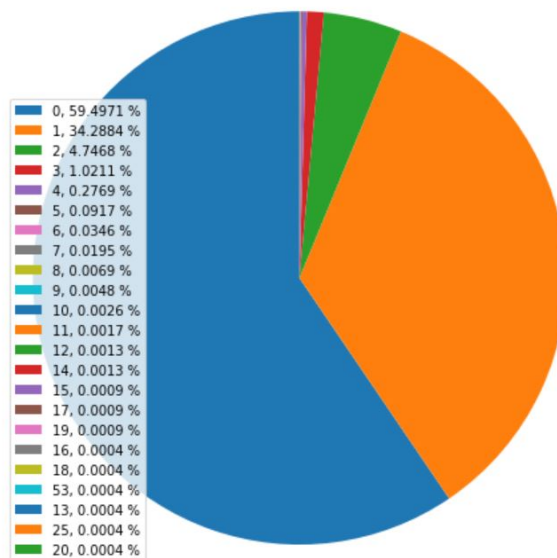
Figuur 8.

Zoals op figuur 8 te zien is, is het grootste deel van de incidenten van het type gewapende overval. Ook familie en vrienden zijn vaak de oorzaak van wapengeweld. Mass shootings hebben het kleinste aandeel van alle typen incidenten.

4. Verhouding tussen aantal gewonden en aantal doden

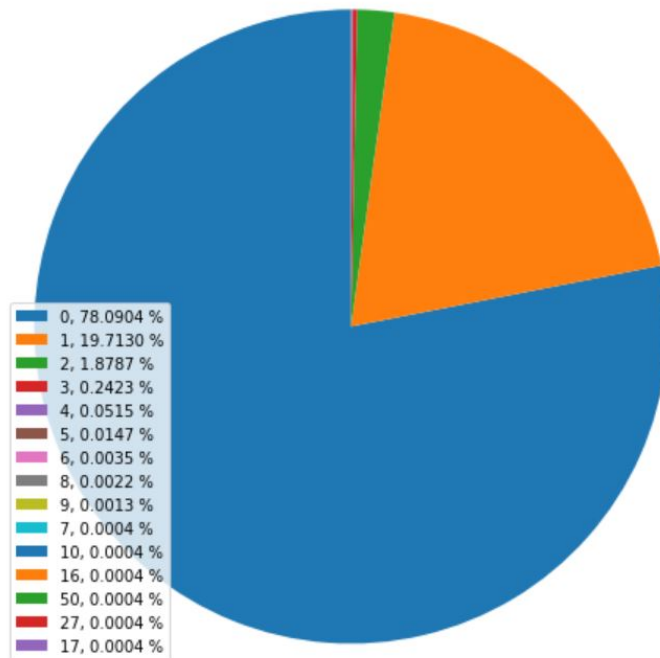
In onderstaande pie charts (fig. 9,10) zijn respectievelijk alle voorkomens van aantallen gewonden per incident en aantallen doden per incident weergegeven met frequenties en percentages.

Amount of people injured per incident



Figuur 9.

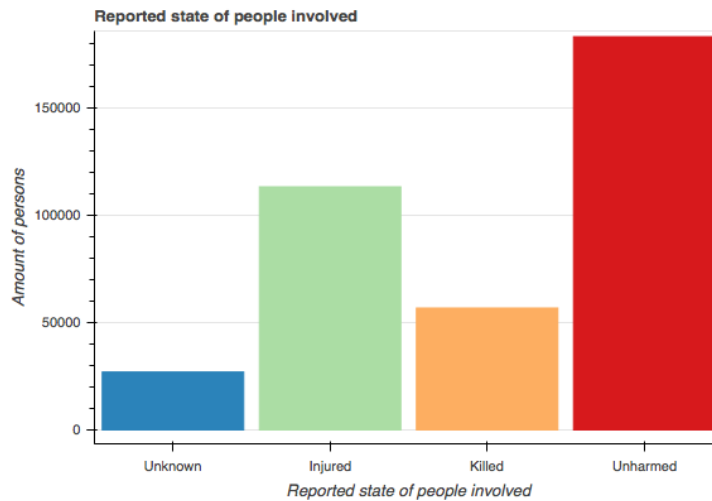
Amount of people killed per incident



Figuur 10.

Uit beide pie charts blijkt dat er relatief weinig incidenten voorkomen waar meer dan twee gewonden of meer dan twee doden vallen. Incidenten met nul doden en incidenten met nul gewonden zijn het meest frequent.

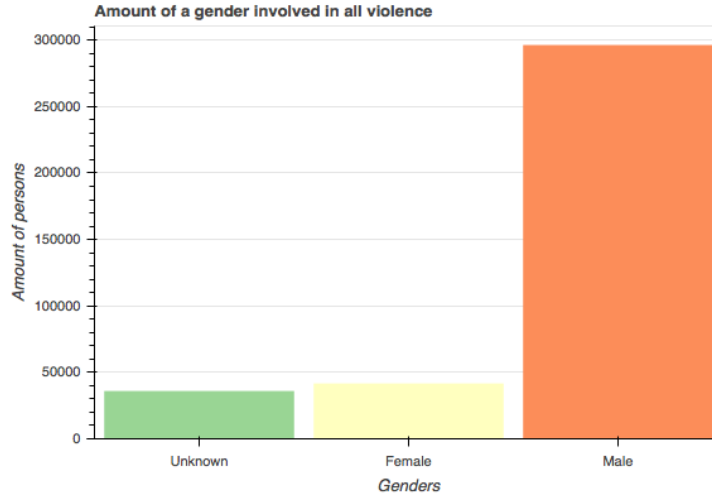
De volgende bar chart (fig. 11) toont het aantal doden en gewonden in verhouding tot het aantal ongedeerden. Het aantal ongedeerden is groter dan het aantal doden of gewonden, zoals de pie charts ook toonden, maar het totale aantal slachtoffers is groter dan het aantal ongedeerden.



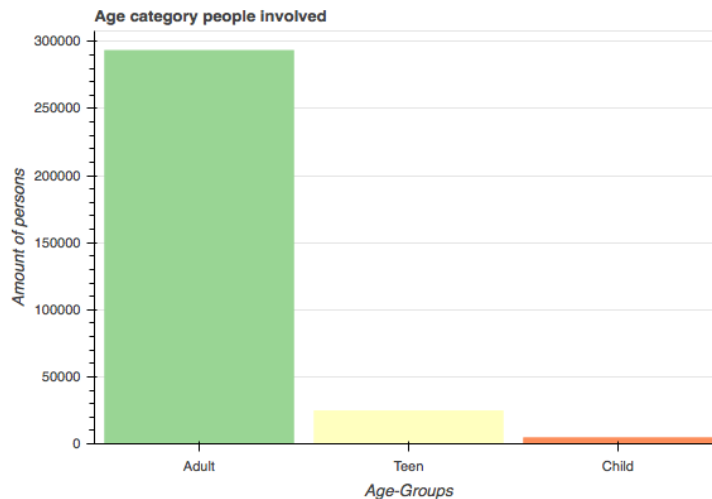
Figuur 11.

5. Verhoudingen in geslacht en leeftijd

Onderstaande bar charts geven respectievelijk de verhoudingen in geslacht (fig. 12) en leeftijd (fig. 13) weer voor de totale verzameling van gerapporteerde incidenten in de afgelopen jaren.



Figuur 12.

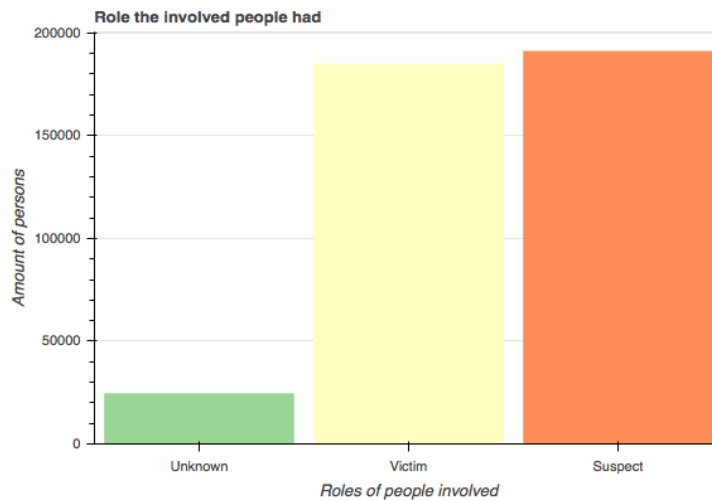


Figuur 13.

Uit de tabellen blijkt dat mannen het meest frequent betrokken raken in wapengeweld. Een relatief groot deel van de betrokkenen is echter van onbekend geslacht. Verder is de grote meerderheid van de betrokkenen volwassen, met op de tweede plaats tieners, gevolgd door kinderen.

6. Verhouding tussen aantal daders en aantal slachtoffers

Onderstaande bar chart (fig. 14) toont van alle incidenten van de afgelopen jaren opgeteld de aantallen slachtoffers en verdachten.



Figuur 14.

In de totale verzameling van incidenten blijken zich ongeveer evenveel slachtoffers als daders te bevinden. Een klein deel van de betrokkenen in de incidenten heeft een onbekende rol gespeeld.

4 Discussie

Uit de vergelijking van het aantal incidenten per staat bleek dat Illinois - grotendeels vanwege het aantal incidenten in Chicago - verreweg de andere staten overtrof in wapengeweld. Wat echter wel genoemd moet worden, is dat in Washington D.C., een staat die zo klein is dat hij niet opvalt tussen de absolute waarden, zoveel wapengeweld plaatsvindt dat de staat weggelaten moest worden uit de genormaliseerde visualisatie. Door het kleine aantal inwoners werd de schaalverdeling namelijk zo groot, dat er geen onderscheid meer gemaakt kon worden tussen de rest van de staten. Deze visualisatie geeft dus geen volledig realistisch beeld van het wapengeweld in de VS.

Verder bleken er meer incidenten te zijn aan de kust dan in het binnenland; de voornaamste reden daarvoor is dat er aan de kust meer grote steden zijn - de genormaliseerde waarden zijn vergelijkbaar met de rest van de VS. Het feit dat in Chicago buitengewoon veel incidenten zijn gerapporteerd is deels doordat het een relatief grote stad is, maar dan nog steekt Chicago ver uit boven steden als New York, dat buiten de top tien valt.

Een zorgwekkende bevinding was dat het wapengeweld sinds 2014 jaarlijks is toegenomen en dat daar geen expliciete verklaring voor te vinden is. Wel blijkt uit onderzoek een sterke correlatie tussen het aantal wapens in een samenleving en de hoeveelheid wapengeweld (Lopez, 2018); een toename aan wapens in de VS is dus een mogelijke oorzaak.

Voor de jaarlijkse piek in de zomermaanden is wel een verklaring gevonden: het verband tussen temperatuur en misdaad is al eerder opgemerkt en onderzocht en daaruit bleek een duidelijke positieve correlatie (Dahl, 2012). Uit datzelfde onderzoek bleek dat hogere temperaturen leiden tot algemene irritatie, met als gevolg een toename in gewelddadige misdaad (in tegenstelling tot geweldloze misdaad). Dit verklaart ook dat de piek in 2017 een aantal maanden vroeger begon: in dat jaar werd er in de gehele VS een gemiddeld hogere temperatuur gemeten (NOAA, 2018). Waarschijnlijk is dus eerder in het jaar de temperatuur bereikt waarop de irritatie het hoogst was.

Uit het bestuderen van de verschillende incidenten bleek dat er in totaal opvallend weinig doden vallen: met de aandacht die in het nieuws wordt gegeven aan mass shootings was de verwachting een hoger aantal doden te vinden. Dit wordt echter niet ondersteund door figuur 8 (relaties tussen betrokkenen): mass shootings hebben het kleinste aandeel en dus amper invloed.

Mannen en volwassenen zijn verantwoordelijk voor de meerderheid van de incidenten; dit was geen onverwachte bevinding, hoogstens het feit dat tieners een vrij klein aandeel hebben verbaasde ons.

Wat wel opviel was het feit dat het totale aantal slachtoffers van de afgelopen jaren kleiner is dan het aantal daders. Door de buitenproportionele aandacht voor mass shootings, waar één of een klein aantal personen vele slachtoffers maakt, werd een veel groter aantal slachtoffers verwacht. Uit nader literatuuronderzoek bleek echter dat tweederde van de dodelijke gevallen van schietincidenten zelfmoordgevallen zijn (Lopez, 2018). Aangezien een zelfmoord één dader en één slachtoffer geeft, leidt een groot aandeel van deze gevallen (in combinatie

met een relatief klein aantal mass shootings) tot een gelijke suspect/victim-verdeling. Bovendien stijgt het aantal daders door het grote aantal gewapende overvallen onder de incidenten: deze worden niet zelden uitgevoerd met handlangers. Daar komt het grote aantal ongedeerden bij incidenten bij, wat leidt tot een groter aantal daders dan slachtoffers.

Een mogelijk onderwerp voor vervolgonderzoek is het verband tussen rijkdom en wapengeweld: worden er meer incidenten gerapporteerd in staten waar armoede heerst? Verder kunnen er ook met deze dataset andere verbanden worden onderzocht, mits deze aangevuld wordt. Van een aantal incidenten was bijvoorbeeld het aantal betrokken wapens bekend, maar dat waren te weinig gevallen om mee te kunnen nemen in het onderzoek. Deel van het oorspronkelijke plan was om een aantal scatterplots te maken om meerdere verbanden tegelijk te onderzoeken, maar daar was uiteindelijk te weinig tijd voor. Bovendien is er een poging gedaan om regressie toe te passen op de data, maar de dataset bleek hier niet geschikt voor te zijn. Hetzelfde gold voor clustering-technieken. Ten slotte zou het voor een vervolgonderzoek interessant zijn om meer multivariate visualisaties toe te passen, waar ze nu vooral univariate zijn.

References

Igielnik, R. (2017, 22 juni). Key takeaways on Americans' views of guns and gun ownership.

Geraadpleegd op 28 juni 2018, van <http://www.pewresearch.org/fact-tank/2017/06/22/key-takeaways-on-americans-views-of-guns-and-gun-ownership/>

Dahl, J. (2012, 6 juli). Hot and bothered: Experts say violent crime rises with the heat.

Geraadpleegd op 28 juni 2018, van <https://www.cbsnews.com/news/hot-and-bothered-experts-say-violent-crime-rises-with-the-heat/>

Lopez, G. (2018, 18 mei). America's unique gun violence problem, explained in 17 maps and charts.

Geraadpleegd op 28 juni 2018, van <https://www.vox.com/policy-and-politics/2017/10/2/16399418/us-gun-violence-statistics-maps-charts>

NOAA National Centers for Environmental Information, State of the Climate: National Climate Report for Annual 2017, published online January 2018.

Geraadpleegd op 28 juni 2018, van <https://www.ncdc.noaa.gov/sotc/national/201713>.