Technisch Rapport: Analyse van het wapenmisbruik in De Verenigde Staten

Tessel van Roozendaal, Jeroen van Wely, Raymon van Dijk, Chang Jiang

Data Analysis and Visualization

June 2018

1 Inleiding

De bedoeling van dit project is om het wapen-misbruik in De Verenigde Staten in de periode van 2013 tot maart 2018 te analyseren. Hierbij zullen drie vragen beantwoord worden. Ten eerste zal het verschil tussen staten, steden en jaren worden onderzocht en duidelijk worden weergegeven met behulp van grafische en niet-grafische middelen. Ten tweede zijn de opvallende patronen en verassende vondsten uitgewerkt. Ook dit is op verschillende manieren zo goed mogelijk geprobeerd weer te geven. Ten derde is de relatie tussen de schutter en het slachtoffer door de jaren heen onderzocht. Als laatste zijn aanvullende datasets met betrekking tot armoede gebruikt om opvallende ontdekkingen te kunnen verklaren, en de analyse verder uit te breiden.

De verwachtingen voorafgaand aan het onderzoek waren vrij eenduidig. Gedacht werd dat het wapengeweld met de jaren toe zou nemen, waarbij de Zuidelijke staten en de staten aan de Oostkust de centra zijn van het geweld. Deze gedachten werden ondersteund door het nieuws rondom wapen-misbruik van de afgelopen jaren. De relatie tussen de schutter en zijn slachtoffer(s) viel nog te betwisten. Er vallen steeds meer doden door zogehete 'mass-shootings' [1]. Aan de andere kant neemt ook het geweld tussen gangs steeds meer toe [2]. Verder blijft het aantal familie- en partner zaken ook stijgen. Er werd daarom verwacht dat deze drie types relaties tussen de schutter en slachtoffer(s) ongeveer even vaak zouden voorkomen.

Daarnaast werd verwacht dat er een correlatie bestaat tussen het armoedecijfer van een staat en de hoeveelheid incidenten die er plaatsvinden. Er is immers veel onderzoek gedaan naar de invloed van armoede op criminaliteit. De vooropstaande conclusie van dit soort onderzoeken is altijd dat er sprake is van meer criminaliteit wanneer er sprake is van meer inkomensongelijkheid en armoede binnen een gebied [3].

2 Methode

In dit hoofdstuk zullen de stappen worden omgeschreven die zijn genomen om tot ons uiteindelijke resultaat te komen. Hierbij is gefocust op de grootste keuzes en overwegingen die zijn gemaakt. Dit om niet in de allerkleinste details te blijven hangen, en het enigszins behapbaar te maken voor de lezer. Aangezien het analyseren van de data leidde tot heel veel verschillende algoritmes en programma's, zal met name de procedure van het project beschreven worden. Hierin worden de gemaakte keuzes uitgebreid toegelicht.

2.1 Procedure

In deze sectie zullen de verschillende onderdelen van het project worden toegelicht. Hierbij wordt in het bijzonder aandacht gegeven aan de keuzes die tijdens het project zijn gemaakt.

2.1.1 Data Cleaning

Het eerste onderdeel van het project bestond uit het opschonen van de dataset. Dit is nodig om te doen omdat data sets geregeld missende gegevens bevatten, of in een bepaalde vorm aangeleverd worden die niet gemakkelijk kunnen worden gebruikt tijdens het analyseren van de data. Het is daarom van belang om de dataset kritisch te bekijken, eventuele missende data aan te vullen, en de data in een dusdanige vorm te krijgen dat het gebruikt kan worden voor analyse.

De voor dit project gebruikte dataset bevatte een hoop kolommen met weinig tot geen data. Bij de hoorcolleges is verteld dat als meer dan 40 procent van een kolom leeg is, deze kolom niet gebruikt mag worden. Bij deze kolommen was dit het geval, dus er is er daarom dan ook voor gekozen om ze te verwijderen uit de dataset. Daarnaast bevatten een hoop van de kolommen data die niet nuttig bleek om te analyseren. Het ging hierbij bijvoorbeeld om de namen van de daders, de url van de website waarop een artikel over het incident staat, aantekeningen van de politie en het id-nummer van het incident. Deze kolommen bevaten voor ieder incident verschillende data. Het is daarom zo goed als niet mogelijk om dit op een bepaalde manier met elkaar te vergelijken of te analyseren. Ook bij deze kolommen is ervoor gekozen om ze uit de dataset te verwijderen.

De dataset bevat informatie over de locatie van het incidenten in verschillende vormen: de staat van het incident, de lengte- en breedtegraad waar het incident plaatsvond, de stad of county van het incident en als laatste het adres. Bij veel incidenten ontbrak echter informatie over het adres van het incident. Om een goede analyse te kunnen maken van de gevaarlijkheid van de steden is het nodig om van ieder incident het adres, en in het bijzonder de stad, te weten. Aangezien deze informatie bij veel incidenten mist, is besloten om een reversed-geocode, gebruikmakend van een Google API, te bouwen. Hierbij

is het de bedoeling dat de lengte- en breedtegraden van de incidenten worden gebruikt waarbij het adres mist om hier een bijbehorend adres aan te koppelen. Er hier een stuk code voor geschreven en dit leek te werken. Het gaf echter als output de adressen in een ander soort format dan er in de rest van de dataset werd gebruikt voor het noteren van het adres. Vervolgens is er gekozen om alle adressen te laten aanpassen door de reversed-geocoder, zodat alle adressen in dezelfde vorm zouden staan, dit is immers van belang om het op een eenduidige manier te kunnen analyseren. Helaas bleek dat de Google API het niet toestond om meer dan duizend adressen per dag op te vragen. Uiteindelijk is er toen voor gekozen om het aanvullen van de adressen achterwege te laten, en de focus te verplaatsen naar het direct gebruiken van de lengte- en breedtegraad.

Tenslotte waren veel cellen in de dataset leeg. Aangezien dit problemen oplevert tijdens het analyseren, is er een kort stukje code geschreven om alle lege cellen te vervangen door 'NA'. Na deze stappen uitgevoerd te hebben, was de dataset voldoende opgeschoond om ermee aan het werk te gaan.

2.1.2 Exploratory Data Analysis

De volgende stap was het uitvoeren van verkennende data analyse. Hierbij is gelet op mogelijk interessante informatie die verder onderzocht kan worden. Er zijn vier verschillende manieren van verkennende data analyse:

- Univariate non-graphical: verdeling van het sample en eerste conclusies
- Univariate graphical: focus op bepaalde variabele voor groep van n samples, kwalitatieve en subjectieve analyse
- Multivariate non-graphical: relatie tussen twee of meer variabelen
- Multivariate graphical: grafisch weergeven van relatie tussen twee of meer variabelen

Alle vier deze manieren van analyse zijn aan bod gekomen tijdens de eerste verkenning van de data. Dit heeft veel inzicht gegeven in de data, waarna er een aantal gegevens zijn gekozen om verder uit te lichten.

Om de eerder genoemde vragen naar tevredenheid te kunnen beantwoorden is er gekozen om de verschillen tussen staten, steden en jaren uit te lichten. Hierbij wordt gebruik gemaakt van univariate graphical analyse. Daarnaast is de relaties tussen schutter en slachtoffer verder uitgezocht, met hierbij de focus op het vaker voorkomen van gang gerelateerde of familie gerelateerde incidenten. Dit is gedaan door een multivariate graphical analyse uit te voeren. Als laatste zijn de opvallende gegevens uitgelicht. Dit is met name gedaan door een scatterplot te maken over de kaart van Amerika heen, om op deze manier duidelijk weer te geven op welke plekken in het land het geweld gecentreerd is. De reden voor deze centrering is op verschillende manieren geprobeerd te verklaren, dit zal in het onderdeel resultaten verder worden besproken.

2.2 Belangrijkste algoritmes

De grafieken zijn verschillende implementaties van bokeh met allen verschillende functies en opmaak. In het begin zijn de wat simpelere grafieken gemaakt zoals de tabel over het aantal doden en gewonden per maand, waarbij de zes jaren bij elkaar zijn opgeteld. Zodra er door het maken van deze ietwat simpelere grafieken en tabellen handigheid was verkregen in het gebruiken van bokeh, zijn ingewikkeldere interactieve grafieken gemaakt die meer inzicht bieden in de data.

Het creëren van de kaart van Amerika met hierover een scatterplot met de gegevens van alle incidenten was het meest uitdagende deel van dit project. Het is de bedoeling dat deze kaart een compleet beeld biedt van de locatie van de incidenten, de gradatie van gevaarlijkheid van de staten, en specifieke informatie biedt per incident. Om de kaart te maken is gebruik gemaakt van de verschillende libraries die bokeh biedt, waaronder de mogelijkheid om de staten en counties van Amerika als achtergrond te gebruiken van een grafiek. Uit deze kaart zijn echter Hawaii en Alaska verwijderd omdat deze twee staten een stuk verder weg liggen, en de kaart dusdanig van vorm veranderde dat hij minder bruikbaar was. Verder is er een hovertool gemaakt die per incident informatie biedt over de datum van het incident, de staat waarin het gebeurde, de stad of county, het adres, het aantal gewonden en het aantal doden. Om dit te doen zijn de benodigde kolommen uit de dataset in de door bokeh te gebruiken vorm gegoten door gebruik te maken van de functie ColumnDataSource.

Deze kaart biedt echter alleen een absoluut beeld van het aantal incidenten en dus de gevaarlijkheid van de staten. Er is hierbij geen rekening gehouden met de populatiegrootte van een staat. Daarom is er besloten om nog een kaart toe te voegen waarin de relatieve gevaarlijkheid van de staten wordt weergegeven. Het spreekt immers voor zich dat de kans groter wordt dat er incidenten plaatsvinden naarmate er meer mensen in een bepaald gebied wonen. Om deze kaart te maken is het aantal doden en gewonden per staat gedeeld door de populatiegrootte [4]. Dit gaf het aantal slachtoffers per hoofd van de bevolking. Vervolgens hebben alle staten een score gekregen en aan de hand daarvan worden de kleuren van de staten bepaald. Tijdens het maken van deze kaart bleek dat er in The District of Columbia relatief erg veel doden en gewonden vallen. Dit bracht de rest van de kaart volledig uit balans omdat de rest van de staten nu allemaal een gele kleur kregen, en er dus geen duidelijk verschil meer zichtbaar was tussen de rest van de staten. Er is daarom gekozen om The District of Columbia buiten beschouwing te laten bij het toedelen van de scores aan de staten.

2.3 Data-analyse

Om met zekerheid conclusies te kunnen trekken over de data, is met name gekeken naar de variantie en standaardafwijking van de verschillende gegevens. Aangezien veel van de grafieken te maken hebben met de verdeling van de incidenten, is hier geen toets op toegepast. Verder is er alleen een analyse gedaan van feitelijke data, het is daarom niet nodig geweest om foutmarges en significantie van resultaten erbij te betrekken. Dit is immers alleen van belang als er sprake is van enquêtes of uitkomsten die op verschillende manieren geïnterpreteerd kunnen worden. Aangezien er in dit geval alleen analyses worden gedaan van feitelijke informatie, is ervoor gekozen om dit statistische deel achterwege te laten in het onderzoek.

3 Resultaten

In dit hoofdstuk zijn de verkregen grafieken en tabellen te zien.
Er zijn een aantal dingen die meteen in het oog springen wanneer de resultaten worden bekeken. Ten eerste neemt het aantal incidenten per jaar gedurende de jaren vorderen sterk toe, zoals in figuur 1 te zien. Verder is in figuur 2 en 4 te zien dat de piek van het aantal incidenten ieder jaar rond de zomermaanden ligt, er kan dus gesteld worden dat dit de gevaarlijkste maanden van het jaar zijn. Deze trend geldt niet alleen voor het aantal incidenten, ook het aantal ligt hoger in de zomermaanden, dit is te zien in figuur 3.

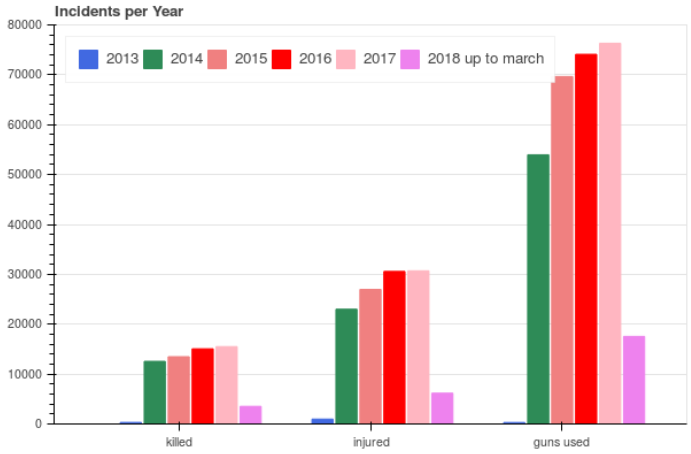

Figure 1: Incidenten per jaar

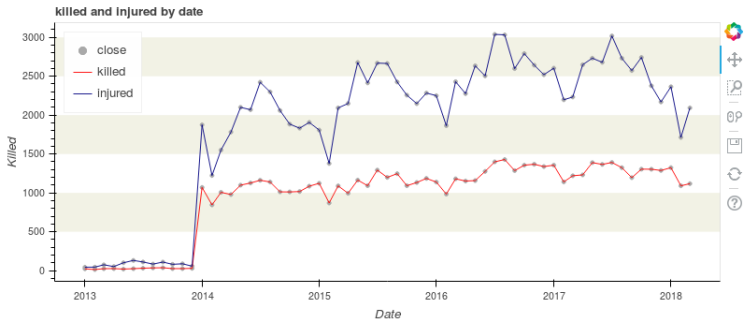

Figure 2: Incidenten 2013-2018

Figure 3: Gemiddelde aantal incidenten per maand voor de periode 2013-2018

#	Month	Killed	Injured	Total
0	Jan	6035	10946	16981
1	Feb	4945	8420	15365
2	Mar	5641	10475	16116
3	Apr	4383	8907	15290
4	May	4830	10244	15077
5	Jun	4886	9797	15130
6	Jul	5276	11259	16535
7	Aug	5127	10809	15936
8	Sep	4779	9766	14535
9	Oct	4791	9756	14547
10	Nov	4848	9087	13935
11	Dec	4927	8936	13863

Figure 4: Totaal aantal incidenten per maand voor de periode 2013-2018

Uit onderstaand cirkeldiagram (figuur 5) blijkt duidelijk dat het bij de meeste incidenten gaat om een gewapende overval, gevolgd door familie- en partner gerelateerde conflicten. In figuur 6 is te zien dat bij het overgrote deel van de incidenten mannen betrokken zijn, zowel als verdachte en als slachtoffer.

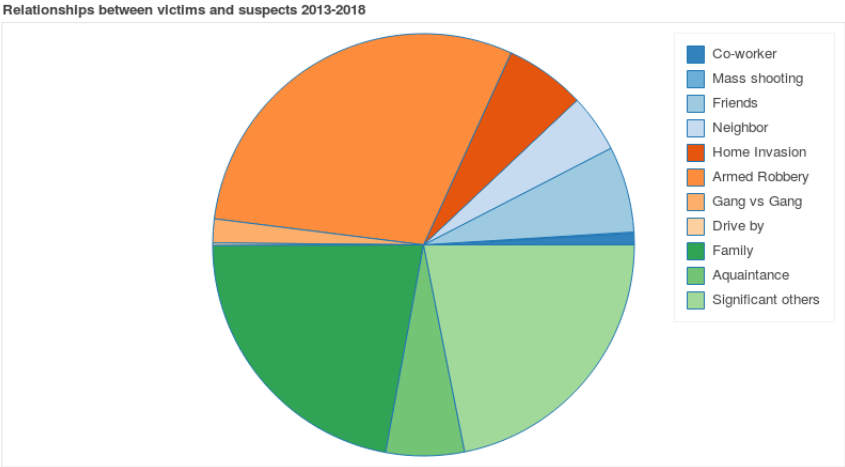

Figure 5: Relatie tussen schutter en slachtoffer

#	Suspect/Victim	Child	Teen	Adult	Gender
0	Suspect	389	10402	139185	Male
1	Victim	2147	9599	122354	Male
2	Suspect	42	542	10820	Female
3	Victim	1276	2232	27603	Female

Figure 6: Verdeling schutter en slachtoffer over man/vrouw en leeftijdsgroep

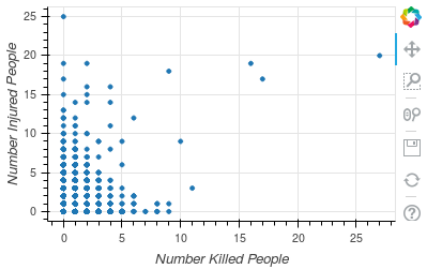

Figure 7: Clustering van hoeveelheid doden en gewonden per incident

In figuur 8 zijn de incidenten verdeeld over de kaart van de Verenigde Staten en de gradatie van de hoeveelheid incidenten van de staten te zien. Hieruit blijkt dat de meeste incidenten plaatsvinden aan de Oostkust, Illinois, Texas, Florida en California. Hierbij wordt echter geen rekening gehouden met de populatiegrootte van de staten, het is dus een absolute analyse. Om ook de relatieve gevaarlijkheid van de staten te kunnen weergeven is figuur 9 gemaakt. Hierin is de relatieve gevaarlijkheid van de staten te zien. Uit deze figuur blijkt dat Illinois, Louisiana, Delaware en The District of Columbia het meest gevaarlijk zijn. Wanneer dit vervolgens wordt vergeleken met figuur 10, waarin het armoedecijfer per staat [5] wordt weergegeven, valt op dat er een zekere overeenkomst is tussen de relatieve gevaarlijkheid en het armoedecijfer van de staat. De resultaten zijn echter niet zo eenduidig, dat hier sluitende conclusies aan verbonden kunnen worden.

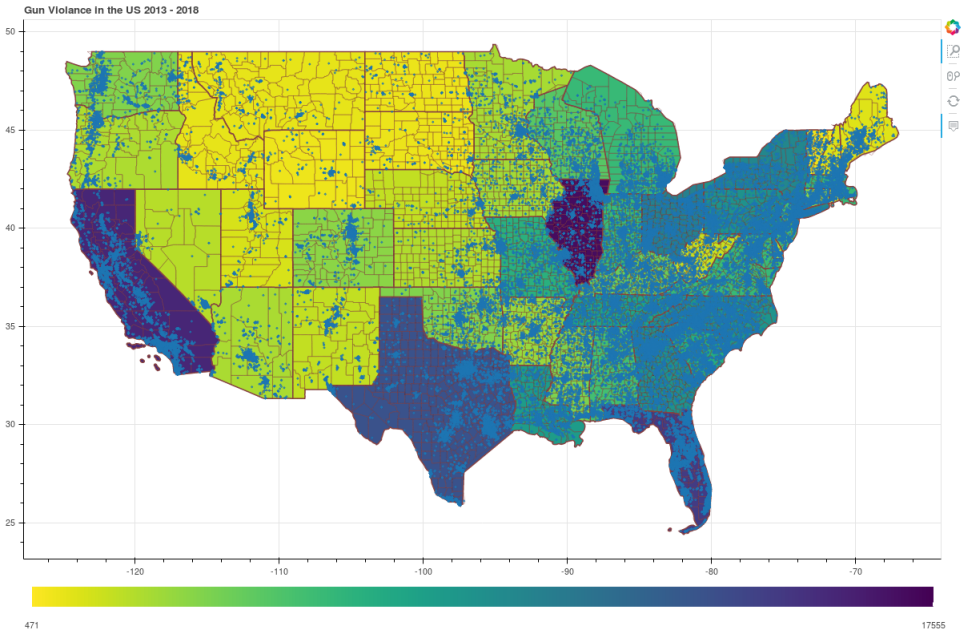

Figure 8: Incidenten als scatterplot op de kaart van de Verenigde Staten

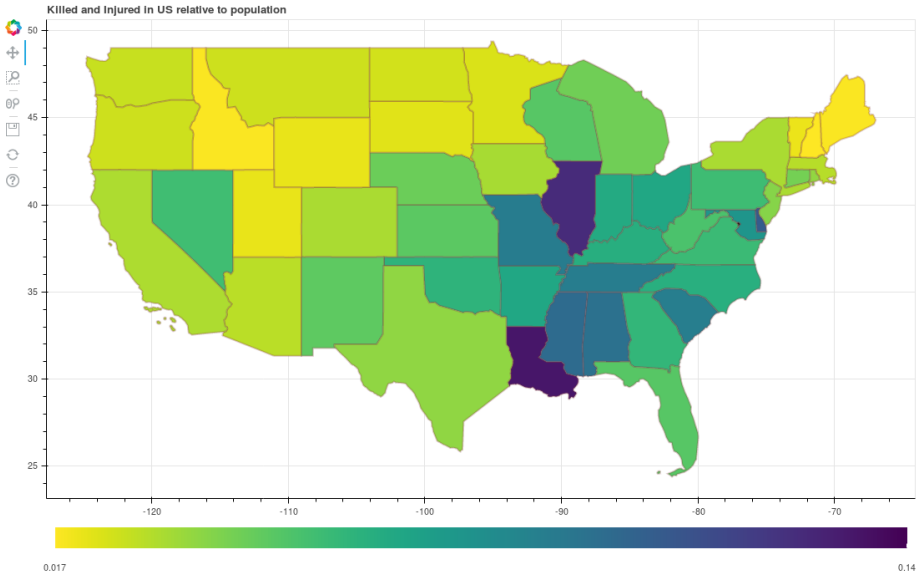

Figure 9: Gradatie van de relatieve gevaarlijkheid van de staten

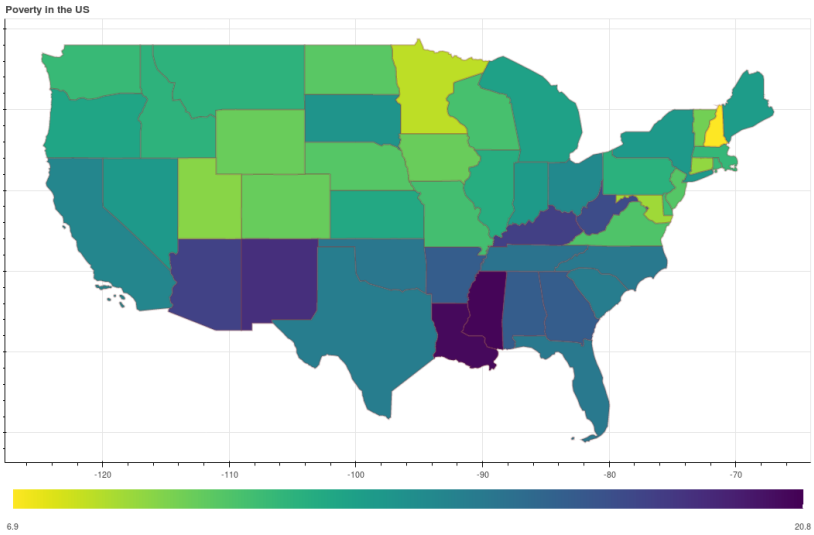

Figure 10: Armoedecijfer van de Verenigde Staten per staat

4 Discussie

Dit onderzoek heeft een aantal belangrijke dingen uitgewezen over het wapen misbruik in de Verenigde Staten. In dit hoofdstuk zullen deze bevindingen toegelicht worden, en er zullen mogelijke verklaringen voor worden gegeven.

Een van de doelen van dit onderzoek was het verschil in incidenten tussen staten, steden en jaren onderzoeken. Hier zijn een aantal dingen over gebleken. Ten eerste is te zien op de kaart van de Verenigde Staten dat California, Illinois en Florida de drie staten zijn waar de meeste incidenten hebben plaats gevonden in de tijdsperiode van 2013 tot maart 2018. Verder is te zien dat de incidenten zich met name centreren aan de twee kusten. In het midden van de Verenigde Staten vinden minder incidenten plaats. Dit geeft echter een enigszins vertekend beeld, omdat er in de kust staten ook een stuk meer mensen wonen dan midden in het land. Wanneer men echter objectief geïnteresseerd is in de meest gevaarlijke staten van de Verenigde Staten kan deze data analyse goed gebruikt worden.

Een andere trend die opviel was de toename van het aantal incidenten gedurende de jaren. Wanneer men kijkt naar figuur 1 is er duidelijk te zien dat er sprake is van een hevige toename vanaf 2013 tot 2017. Van 2018 zijn alleen de gevallen tot maart bekend, maar als de huidige trend wordt doorgerekend zullen er ook dit jaar meer incidenten plaatsvinden dan voorgaande jaren. Bovendien is uit de data gebleken dat de zomermaanden de gevaarlijkste maanden van het jaar zijn, en deze moeten voor 2018 nog komen. Verder valt het op dat er in 2013 opvallend weinig incidenten hebben plaats gevonden. Het verschil is dusdanig groot, dat het vermoeden is ontstaan dat er data mist over dit jaar. Wanneer er echter wordt gekeken naar de datums die voorkomen in dit jaar in de dataset, ontbreken er geen specifieke maanden. Er is daarnaast ook geen informatie beschikbaar over het al dan niet ontbreken van data. Daarom is er aangenomen dat de dataset volledig is. Er is dus sprake van een extreme toename van incidenten vanaf het jaar 2014.

Helaas was de data met betrekking tot de stad waar het incident niet volledig zuiver. De kolom die hier informatie over gaf heeft de titel 'city_or_county'. Deze kolom bevat dus soms de naam van de stad waar het incident plaatsvond, en soms de county. Zoals eerder omschreven in het hoofdstuk 'Methode' is er geprobeerd een reversed-geocoder te bouwen om de volledige adressen van de incidenten te verkrijgen aan de hand van de lengte- en breedtegraad. Dit is echter niet gelukt. Als gevolg hiervan was het niet goed mogelijk om eenduidige informatie te verkrijgen over de stad waarin een specifiek incident plaatsvond. De analyse over de gevaarlijkheid van de steden kon daarom niet uitgevoerd worden. Er is echter wel goed op de kaart te zien waar veel incidenten hebben plaats gevonden. Op die manier kan men toch een enigszins verduidelijkend beeld krijgen van de plaatsen waar de meeste incidenten plaatsvinden.

Met betrekking tot de relatie tussen de schutter en het slachtoffer zijn er een aantal opvallende resultaten naa voren gekomen. Aan het begin van dit onderzoek werd verwacht dat 'Gang vs Gang', 'mass shootings' en familie en

partner zaken de meest voorkomende relaties zouden zijn. Na het analyseren van deze informatie, is gebleken dat de meeste incidenten het gevolg zijn van een gewapende overval. Een verklaring voor het verschil tussen de verwachting en de daadwerkelijke gegevens, is het feit dat de berichtgeving over dit soort gewapende incidenten die de rest van de wereld bereikt met name gefocust is op de wat meer 'schokkende' gevallen. Er wordt relatief weinig gerapporteerd over gewapende overvallen, omdat dit inmiddels bijna tot de orde van de dag hoort in de Verenigde Staten. De tweede en derde meest voorkomende relaties zijn familie en partners. Dit is dus wel in lijn met de verwachting.

Ten slotte is de correlatie tussen het armoedecijfer van een staat en het relatieve aantal incidenten per staat onderzocht. Hieruit kwam enige correlatie naar voren die met name te zien is in de Zuidelijke staten. Over het algemeen waren er echter niet genoeg overeenkomsten te vinden om een duidelijke conclusie te kunnen trekken over het al dan niet aanwezig zijn van een positieve correlatie tussen de twee factoren. Een mogelijke verklaring hiervoor is dat er meer factoren meespelen met de hoeveelheid incidenten, zoals het aanwezig zijn van achterstandswijken en dergelijke. Wanneer er meer factoren bij deze analyse zullen worden betrokken, zal er mogelijk wel een correlatie worden gevonden.

Dit onderzoek zou verbeterd kunnen worden met name door meer factoren en externe data sets erbij te betrekken. Er zal op die manier een completer beeld kunnen worden geschetst van de achterliggende redenen van de incidenten. Wanneer er een duidelijke verband gevonden kan worden tussen een bepaalde externe factor en de hoeveelheid incidenten, zal er mogelijk een beter beleid gevormd kunnen worden om de hoeveelheid incidenten terug te dringen.

References

[1] Federal Bureau of Investigation. *Office of Partner Engagement.* Verkregen van: https://www.fbi.gov/about/partnerships/office-of-partner-engagement/active-shooter-incidents-graphics, juni 2018

[2] National Gang Center. *National Youth Gang Survey Analysis.* Verkregen van http://www.nationalgangcenter.gov/Survey-Analysis, juni 2018.

[3] Kennedy, B. P., Kawachi, I., Prothrow-Stith, D., Lochner, K., Gupta, V. (1998). Social capital, income inequality, and firearm violent crime. *Social science & medicine* 47(1), 7-17

[4] World Population Review. *US States - Ranked by Population 2018.* Verkregen van: http://worldpopulationreview.com/states/, juni 2018

[5] Jessica L. Semega, Kayla R. Fontenot, and Melissa A. Kollar. (2017). *Current Population Reports*, 10-11.