Process book

Willem van der Spek, Silvan Murre, Bram Bakker en Joris Hijstek

Juni 2018

Algemene voortgang

1. 4 Juni 2018
   (a) De dataset *Gun Violence* is gekozen omdat hiervoor interesse was binnen de groep.
   (b) De eerste denkstappen binnen het *cleanen* de dataset zijn gemaakt: er is beredeneerd welke kolommen onnodig zijn en welke datatypes uit de kolomennen geëxtraheerd kunnen worden.
   (c) de onderzoeksvragen zijn tot stand gekomen, zie de README.
2. 7 Juni 2018
   (a) Er is een oplossing gevonden voor de lege cellen in de dataset: de 'fillna' *built-in* van *pandas*.
   (b) Alle onnodige kolommen zijn verwijderd
   (c) De csv is correct geïndexeerd.
3. 8 Juni
   (a) Er is data verzameld buiten de dataset om voor de vuurwapenwetgeving en populatie per staat.
4. 10 Juni
   (a) Er is nog meer data verzameld, ditmaal om de dagen met de meeste incidenten te ordenen. Het visualiseren wordt nog achterwege gelaten.
5. 12 Juni
   (a) De laatste zaken om de dataset op te schonen zijn afgerond. Namelijk het creeëren van een functie die *strings* kan omzetten naar *dictionaries*.

6. 13 Juni

   (a) De eerste data is gevisualiseerd. In plaats van de data voor elk jaar is er voor gekozen om de data per maand over de afgelopen paar jaar te plotten.

   (b) Er is tevens data verzameld over de rol van tieners in massale schietpartijen.

7. 14 Juni 2018

   (a) De file die *strings* omzette naar *dictionaries* bleek niet perfect te werken. De documentatie van de dataset klopte niet voor sommige kolommen. Omwille hiervan is de functie hiervoor aangepast.

   (b) Er is gekozen om met name de *Plotly API* te gebruiken in plaat van *Bokeh*. De reden hiervoor is het grotere gebruiksgemak en betere interactiviteit.

   (c) Een *choropleth map* is gemaakt voor het aantal doden en gewonden per capita. Zie git repository hiervoor.

   (d) Er zijn nieuwe ideeën gekomen om de dataset te visualiseren, namelijk:

      i. De distributie van relaties komen in een cirkeldiagram.

      ii. De doden m.b.t. relaties worden per staat weergegeven.

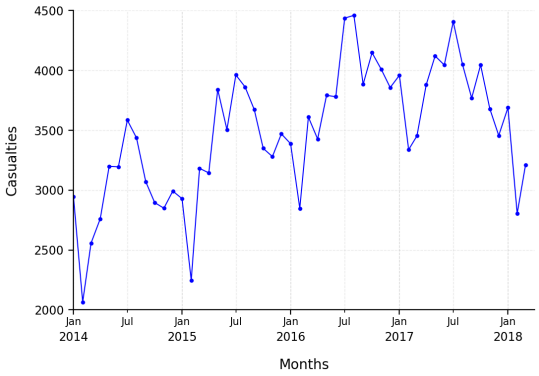
      iii. Er komt een diagram waarin leeftijden relevant zijn voor doden.

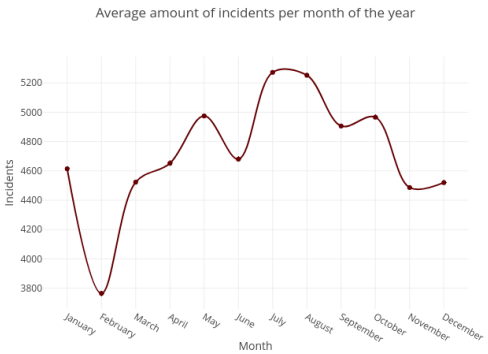
8. 15 Juni 2018

   (a) Het cirkeldiagram voor de relaties is gevisualiseerd.

9. 16 Juni 2018

   (a) Er is beslist dat de data voor elke maan over de jaren heen wordt veranderd naar de data voor het gemiddelde van alle maanden. hieronder de visualisaties naast elkaar ter illustratie.

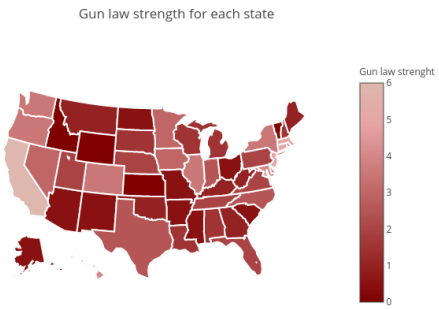


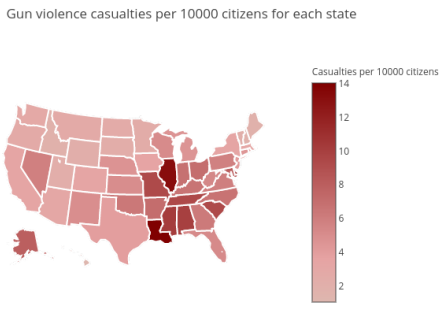


   (b) Aan de hand van een nieuw idee is er een visualisatie gemaakt voor elke dag van de maand.

   (a) De data voor de incidenten verdeeld over de leeftijden is gevisualiseerd.

10. 21 Juni De *choropleth map* voor de vuurwapenwetgeving per staat is toegevoegd, hieronder de visualisatie naast degene van aantal doden en gewonden per capita ter ondersteuning.





11. Er is veel verwarring over de *In-depth analysis*, het is namelijk onduidelijk hoe dit proper kan worden toegepast op onze dataset. De enige mogelijkheid die wij tot nu toe hebben bedacht is een regressielijn voor een puntengrafiek van de meest continue variabelen.

12. 23 Juni

    (a) In verband met het gebrek aan mogelijkheid tot *In-depth analysis* is de voortgang van het project enigzins tot stilstand gekomen.

13. 25 Juni

    (a) De eerste pogingen tot het maken van een website zijn gemaakt.

    (b) Het *embedden* van de plots blijkt problemen te geven voor de grootte van de plots. Een oplossing volgt.

    (c) De scatterplot met haar regressielijn zijn gemaakt. Echter Plotly lijkt niet flexibel genoeg om de regressielijn voor de punten te plotten. Er wordt nagedacht over een oplossing hiervoor.

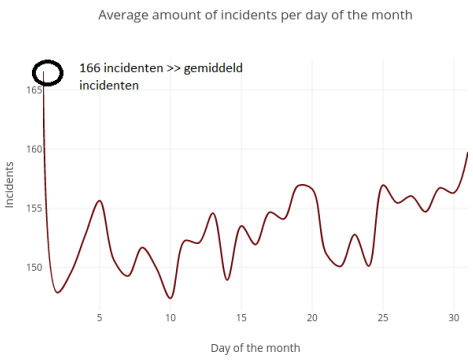
14. 26 Juni

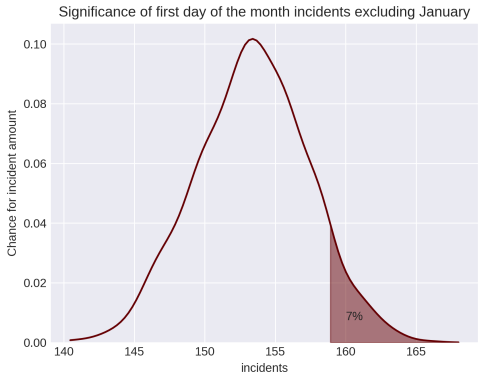
    (a) Kleine wijziginen zijn gemaakt in de leeftijdsdistributie. Hier zijn namelijk percentages gebruikt in plaats van harde aantallen.

15. 27 Juni

    (a) De layout voor de website is neergezet. Het idee om de visualisaties te laten coheren is tot nu toe om ze per onderzoeksvraag in te delen.

    (b) Na een ingeving is er toch animo voor *In-depth analysis*. Het idee is namelijk om de significante waarde bij het significant aantal incidenten voor de eerste dag van de maand te onderzoeken en hierop een statistische toets uit te voeren.





    (c) Blijkbaar bewerkstelligt de eerste van januari deze extreem hoge waarden, dit bevestigt ons vermoeden!

16. 28 Juni

    (a) De logische opbouw van de visualisaties is herzien. In plaats van het ordenen van de visualisaties per vraag, zijn de visualisaties geordend op eigenschappen. De eigenschappen die we hiervoor gebruiken zijn:

        i. tijd
        ii. staten
        iii. leeftijd
        iv. relaties
        v. extra's

    (b) De website is verder qua stijl verbeterd.

    (c) De *git repository* is aangepast aan de hand van de parameters die zijn meegegeven op *UvA Cannvas* zijn gehandhaafd ter ordening. Dit gaf behoorlijjk wat problemen met de html en css bestanden, niettemin is na wat geploeter het probleem opgelost om alle relatieve paden te herschrijven.

Logboeken

Bram Bakker

4 juni + 5 juni: data set uitkiezen, testen connecten met github, pushen en pullen en planning.
6 juni: geprobeerd om met csv een goede 'to dict' functie te maken wat uiteindelijk niet gelukt was. - 2 uur
8 juni: Voor het onderzoeken naar het aandeel van massale schietincidenten door jongeren begonnen met werken aan een functie die de informatie over de leeftijd van een betrokkene kan matchen met of de betrekonne een suspect of victim is. - 2 uur
12 juni:Code geschreven die cijfers met elkaar kan vergelijken om een beeld te krijgen van de data over jongeren - 1 uur
13 juni:Door het gebruik van to dict code ontwikkeld door groepje, die strings naar dictionaries om kan zetten kon uiteindelijk een werkende functie gemaakt worden die de verdeling van mass killings en shooting incidents weergeeft - 3 uur
14 juni: probleem tegengekomen in exploratory data waardoor de cijfers onjuist blijken te zijn, dacht het verholpen te hebben - 1 uur
15 juni: Code voor schietincidenten jongeren volledig geoptimized, veel nutteloze code verwijdert - 3 uur
18 juni: Data correct gevisualized in matplotlib - 1 uur
19 juni: teen shootings uitgebreid naar demographics met mannen en vrouwen en naar interessante datapunten gezocht - 5 uur
21 juni: De visualization omgezet van matplotlib naar plotly wat nog een werk was om werkend te krijgen op een windows laptop waar eerst anaconda op gezet moest worden - 3 uur
22 juni: Eerste opzet inleiding geschreven - 1 uur
23 juni: alle demographics veranderd van ruwe cijfers naar percentages en uitgebreid naar meerdere onderzochte types. -2 uur
24 juni: Inleiding verbeteren en methode geschreven. - 3 uur
26 juni: Meerdere onderzochte types weggehaald aangezien ze niet significant speciaal waren en gepushed naar git - 10 minuten
27 juni: technisch rapport uitbreiden en verbeteren - 3 uur
28 juni: Website helpen ordenen. Technisch rapport finaliseren met groep - 7 uur

Silvan Murre

4 juni: Github repository aangemaakt en de dataset uitgekozen. - 1 uur
5 juni: README Document aangemaakt met daarin de eerste drie vragen en samen met het groepje de hypothesen bedacht en opgeschreven. Ook de csv file uit de dataset gedownload en in eigen repository gezet. Ten slotte nog geholpen met het verwijderen van kolommen uit het csv bestand die we voor de analyse niet nodig hadden. - 2 uur
6 juni: Plannen gemaakt voor het toevoegen van nieuwe data: Gun Law Strength per staat, deze kunnen vergeleken worden met het aantal doden en gewonden per staat. Alvast besloten hoe voor elke staat een score kan worden toegekend aan de hand van hoe sterk de wapenwetgeving/hantering is voor die staat. - 3 uur
7 juni: Een dictionary gemaakt met voor elke staat een score van de Gun Law Strength. Vervolgens samen met Willem een dictionary voor de populatie per staat gemaakt. Beide dictionaries hebben we ten slotte in het csv bestand verwerkt. - 2,5 uur
12 juni: Plannen gemaakt en info opgezocht voor het maken van een kaart-plot die het aantal doden en gewonden per 10.000 inwoners per staat weergeeft. Hetzelfde gedaan voor een kaart-plot (chloropleth) die de Gun Law Strength per staat weergeeft. - 2 uur
13 juni: Code gegschreven die data uit het csv bestand haalt en vervolgens hier een dictionary van gemaakt. De dictionary bevat het aantal doden en gewonden per 10.000 inwoners per staat. - 3 uur
14 juni: Erachter gekomen dat het niet goed lukte om de data uit de dictionary in de kaart-plot van USA te zetten. Hierna bezig geweest met het schrijven van code die ervoor zorgt dat er een apart csv bestand gemaakt. Het csv bestand moet de gegevens van de 51 staten bevatten die nodig zijn voor het maken van de kaart-plot. - 3,5 uur
15 juni: Code geschreven die ervoor zorgt dat de twee kaart-plots gemaakt wordt met de data uit het csv bestand (casualtiesper10000citizens.csv). De plots werden alleen nog niet gemaakt door verschillende errors. - 2 uur
19 juni: Bug van 15 juni gefixt; voor het maken van de twee kaart-plots waren de 2-letterige codes per staat nodig. Hiervoor heb ik een dictionary gemaakt en deze in het csv bestand uitgeschreven. Ook de kleuren van de plots aangepast. - 2,5 uur
20 juni: Extra code geschreven. Van het totale aantal casualties is per staat het percentage doden en het percentage gewonden berekend, vervolgens is dit in de beschrijving bij elke staat toegevoegd. - 2,5 uur
21 juni: Code geschreven die het aantal decimalen van de getallen in de beschrijving van elke staat limiteert naar twee decimalen. Ook code geschreven die ervoor zorgt dat de populatie per staat in miljoenen wordt weergegeven - 2 uur
26 juni: Basic indeling voor de website gemaakt (Banner en vragen). - 1,5 uur
27 juni: Aantal plot-visualisaties in de website verwerkt. - 3 uur
28 juni: 11:00-18:30: Gehele website afgemaakt en meegeholpen met technisch rapport. 21:00-23:00: Website-info toegevoegd.

Joris Hijstek

4-7 juni: Dataset uitgekozen (Gun violence), basis opzet gedaan met github, taakverdeling en planning. - 1 uur
8-10 juni: Gewerkt aan basiskennis van Pandas, bijv. hoe te zien of een bepaalde cell geen informatie bevat, pre-processing van dataset - 3 uur
11-15 juni: Eerste plot opgezet met alle staten van Amerika en de meest voorkomende relaties bij moorden. Dit koste wat meer tijd vanwege het feit dat ik variabelen met strings moest kunnen matchen. Dit heb ik uiteindelijk opgelost d.m.v. dictionaries voor elke oorzaak. - 6 uur
16-20 juni: Tijdelijke webserver opgezet. Ik liep hier tegen problemen met poorten aan, en het editen op afstand. (Dit was voordat we wisten dat github ook een html-optie aanbiedt. - 9 uur
21-25 juni: Bezig geweest met het visuele aspect van de website, vooral met het goed opstellen van de verschillende grafieken. HTML en CSS besloten vervelend te doen, dus heb ik uiteindelijk bij stackoverflow om hulp gevraagd. (https://stackoverflow.com/questions/51010402/html-css-plotly-plot-size/51019530) - 8 uur
26-27 juni: Vooral gewerkt aan het uiterlijk van de grafieken, zoals het kleurenpalet, en de layout van de as-labels. - 2 uur
28 juni: Gewerkt aan verslag, website klaar voor gebruik gemaakt (op github), code voor grafiek sneller gemaakt en een fatale bug gefixt die de data verkeerd weergaf. 7 uur
fdsafasd

Willem van der Spek

04-06-2018: Hoorcollege bijgewoond en onderhandeld met dhr. de Wolff en andere groepjes om in een normaal groepje te komen. - 4.5 uur

05-06-2018: Groffe lijnen uitzetten en eerste taken verdelen, verder is er geëxperimenteerd met pandas en zijn onnodige kolommen verwijderd. Bijeenkomst met begeleider bijgewoond. - 4.5 uur

06-06-2018: Ordenen van repository en kennis opdoen over het opschonen van data. - 2 uur

07-06-2018: script voor data per maand toegevoegd, script toegevoegd om voor meedere csvs te cleanen en data toegevoegd aan de dataset m.b.t. wapenwetgeving en populatie. Bijeenkomst begeleider bijgewoond - 5 uur

08-06-2018: Script toegevoegd om dagen met de meeste incidenten te vinden verder is de repository verder geordend. - 2.5 uur

10-06-2018: Verder gewerkt aan script met meeste incidenten, ingelezen in git bash documentatie. - 2.5 uur

11-06-2018: Hoorcollege gevolgd - 2 uur

12-06-2018: Script geschreven om data per maand te visualizeren, geëxperimenteerd met matplotlib en script geschreven om strings uit csv naar dictionary te transleren. Bijeenkomst met begeleider verder bijgewoond. - 8.5 uur

13-06-2018: Script om string naar dictionary te converteren geperfectioniseerd en het team uitleg gegeven. Distributie gemaakt van relaties voor incidenten. - 4 uur

14-06-2018: Team geholpen met schrijven van scriptjes, bijeenkomst met begeleider bijgewoond. - 3 uur

15-06-2018: Ingelezen in Plot.ly en geëxperimenteerd. Visualisatie voor data per jaar toegevoegd - 4 uur

16-06-2018: Visualisatie voor data per maand opnieuw ontworpen, Visualisatie voor data per dag van de maand toegevoegd. - 5 uur

18-06-2018: Hoorcollege gevolgd. - 2 uur

19-06-2018: Bijeenkomst met team bijgewoond, gewerkt aan wat visualisaties en gespeculeerd over verschillende analysetechnieken. - 4 uur

20-06-2018: Beginselen van de website gemaakt, regressietechnieken toegepast, TA meeting. - 3.5 uur

24-06-2018: Gewerkt aan het visualiseren van schietincidenten per staat. 1 uur

26-06-2018: TA meeting, verder gewerkt aan de opmaak van de visualisatie van de regressie en het team geholpen. Incidenten met meeste doden in top 10 lijst verwerkt. 4 uur

27-06-2018: Gewerkt aan het verslag, Statistische analyse uitgevoerd op de data over dagen van de maand. - 5 uur

28-06-2018: Statische analyse gevisualiseerd, structuur van de website bepaald, TA meeting, debuggen van de paden na het herordenen van de repository, top 10 steden met meeste doden berekend, gewerkt aan het verslag en het team veelvuldig geholpen. - 9 uur

29-06-2018: Demo - 4 uur