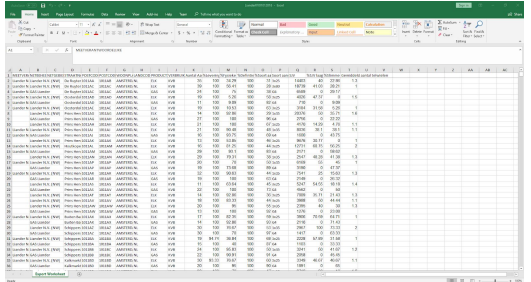

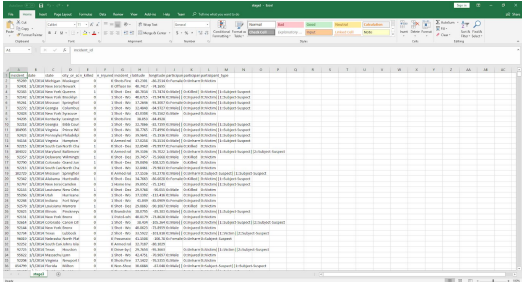


Process book

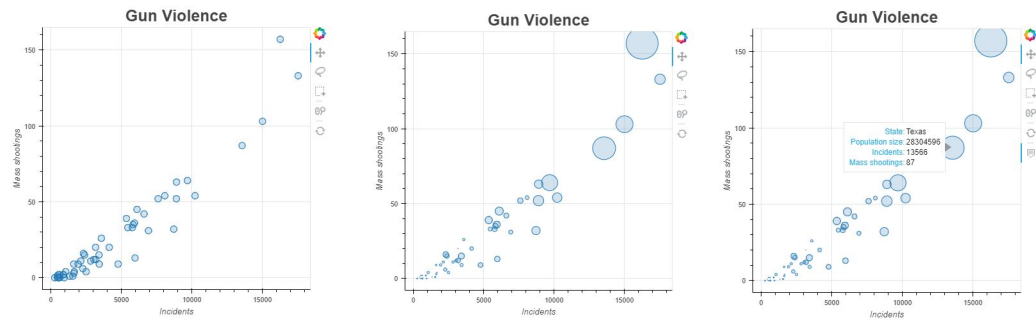
Marijn Alta, Sarah Tol, Hugh Mee Wong en Fengyuan Sun

Week 1	
Maandag	<p>Dataset</p> <p>Bij het opstarten van het project moesten er allereerst keuzes gemaakt worden: welke dataset gaan we gebruiken en hoe gaan we dat aanpakken?</p> <ul style="list-style-type: none">Er is uiteindelijk besloten om de Power Consumption Amsterdam dataset te gebruiken.
Dinsdag	<p>Dataset</p> <p>Vervolgens werd er gekeken naar de mogelijkheden van deze dataset. Kunnen we hiermee al onze (deel)vragen beantwoorden?</p> <p>Deze dataset bleek echter te weinig informatie te bevatten.</p> <ul style="list-style-type: none">Er is unaniem besloten om de Gun Violence dataset te gebruiken. 
Woensdag	<p>GitHub</p> <p>Belangrijk was het opzetten van een GitHub repository: deze was onmisbaar voor de samenwerking.</p> <p>Verder werden er nieuwe deelvragen bedacht en werden de attributen uit onze dataset geclassificeerd. Hierdoor valt er beter te zien welke variabelen zinvol lijken om te analyseren en te plotten. Zo zal hopelijk het project ook efficiënter verlopen.</p> 
Donderdag	<p>Data cleaning</p> <p>Er is simpelweg te veel onnodige of inconsistente data in de dataset. Door deze te <i>cleanen</i> zal er later in het project efficiënter gewerkt worden met de data en worden mogelijke knelpunten voorkomen.</p> <ul style="list-style-type: none">De attribuut <i>age_group</i> is weggelaten, omdat hieruit alleen de groepen tieners en volwassenen te onderscheiden was. Dit was dus te breed.Verder is <i>participant_age</i> weggelaten, omdat deze te incompleet was. 
Vrijdag	<p>Er werd getest of het mogelijk was om met Google API's de incidenten weer te geven op een</p>

	<p>kaart. Dit bleek niet realistisch te zijn, dus was er besloten om dit met de Bokeh library te plotten.</p>
Week 2	
Maandag	<p>Extra bronnen</p> <p>Voor het beantwoorden van de deelvragen was er meer informatie nodig. Daarom werden er extra datasets verzameld: de Politics en de Gun Ownership datasets. Deze werden overigens ook ge-cleaned.</p>
Dinsdag	<p>Exploratory Data Analysis (EDA)</p> <p>Er is deze week begonnen met het in-depth verkennen van de data, waarbij de vier soorten technieken werden toegepast.</p> <div><div><p><i>Univariate non-graphical</i></p><pre>Incidents: 0 : 142456 - 0.5959567880400502 - 59.50567880400502 % 1 : 81972 - 0.3424074453109623 - 34.24074453109624 % 2 : 11457 - 0.047857342762501094 - 4.78573427625011 % 3 : 2461 - 0.01027990927280398 - 1.027990927280398 % 4 : 657 - 0.0027443723657993558 - 0.2744372365799356 % 5 : 215 - 0.000908022310454514 - 0.00090822310454514 % 6 : 80 - 0.000341701510866796 - 0.03341701510866796 % 7 : 45 - 0.00018797070998625726 - 0.018797070998625726 % 8 : 16 - 6.083483021733591e-05 - 0.000683403021733591 % 9 : 11 - 4.594839577441844e-05 - 0.004584839577441844 % 10 : 6 - 2.506276131500967e-05 - 0.00250627613150097 % 11 : 4 - 1.6708507554333977e-05 - 0.001670850755433977 % 12 : 3 - 1.253138065750484e-05 - 0.001253138065750484 % 13 : 1 - 4.17712688583494e-06 - 0.000417712688583494 % 14 : 3 - 1.253138065750484e-05 - 0.001253138065750484 % 15 : 2 - 8.354253777166988e-06 - 0.0008354253777166988 % 16 : 1 - 4.17712688583494e-06 - 0.000417712688583494 % 17 : 2 - 8.354253777166988e-06 - 0.0008354253777166988 % 18 : 1 - 4.17712688583494e-06 - 0.000417712688583494 % 19 : 2 - 8.354253777166988e-06 - 0.0008354253777166988 % 20 : 1 - 4.17712688583494e-06 - 0.000417712688583494 % 25 : 1 - 4.17712688583494e-06 - 0.000417712688583494 % 53 : 1 - 4.17712688583494e-06 - 0.000417712688583494 %</pre></div><div><p><i>Multivariate graphical</i></p></div></div> <p>Univariate EDA werd toegepast om de betrouwbaarheid en consistentie van verscheidene variabelen te onderzoeken, terwijl multivariate EDA gebruikt werd om verbanden en correlaties tussen verschillende variabelen te analyseren.</p>
Woensdag	<p>Visualisatie</p> <p>Belangrijk voor ons eindproduct is de visualisatie van onze data: dit is hetgene dat de aandacht moet trekken tijdens de demo van ons product. Uitgaand van dit principe zijn er ook visuele plotjes gemaakt met de kaart van Amerika.</p> <ul style="list-style-type: none">Er is verder besloten om de plot over politieke voorkeur per staat weg te laten, omdat deze plot niet vergeleken kon worden met de andere data. <div><p>Violent incidents per capita</p></div>

Donderdag

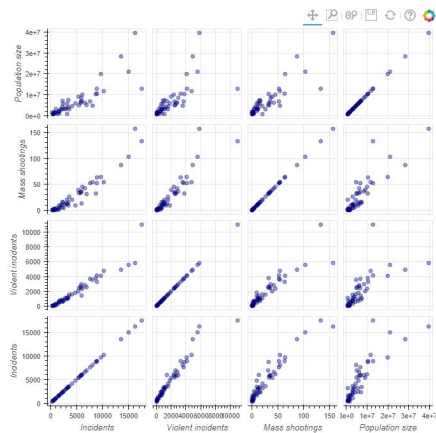
Toevoeging visualisatie



Geleidelijk is er aan de geplote grafieken meer visuele informatie toegevoegd om een beter begrijpbaar grafiek te verkrijgen.

Vrijdag

Scatter plots



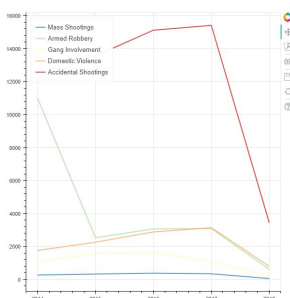
Om effectief data te vergelijken en verbanden te onderzoeken, is er ook gebruik gemaakt van matrix scatter plots. Hiermee werden meerdere variabelen tegelijkertijd geanalyseerd, zodat er een beter begrip van de data verkregen werd.

Week 3

Maandag

Interactief, onderzoek

In de hoorcolleges werd er verteld dat de website, die bij de demo getoond zou worden, echt een verhaal moest vertellen, dus zijn de grafieken interactief gemaakt:



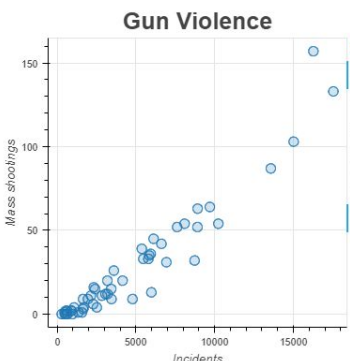
Verder werden lineaire regressie, K-means en clustering op basis van code geïntroduceerd in het hoorcollege en is de stof goed bekeken, opdat er later mee gewerkt kon worden. Het werd

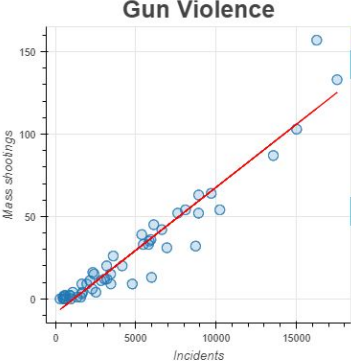
alleen niet meteen begrepen.

Dinsdag

Lineaire regressie & mean squared error

Ondanks dat er gisteren was gekeken naar lineaire regressie, was het nog steeds onduidelijk wat alles precies betekende. Dus is er gekeken naar de voorbeeldcode op Canvas, van Nick de Wolf. Na alles lang met elkaar vergeleken te hebben, was het duidelijk geworden.





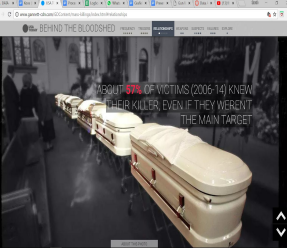
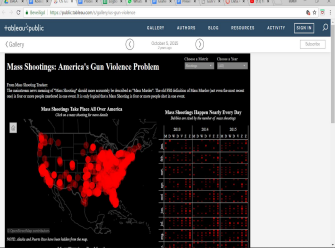
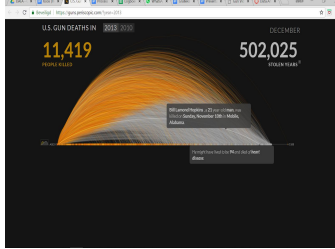
Zonder lineaire regressie

Met lineaire regressie

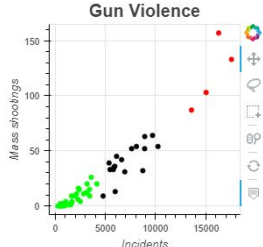
Woensdag

Inspiratie website & K-means

Om te kunnen beslissen welke website design het beste zou zijn, is er gezocht naar inspiratiebronnen: Hieruit zijn de volgende websites gevonden met ongeveer hetzelfde onderwerp.



Na lineair regressie, volgde K-means. Deze was simpel om te gebruiken, via de elleboog techniek. Alleen was het erg lastig om een toepassing voor onze dataset hiervoor te vinden. De K-means gaf namelijk nergens een meerwaarde aan, waardoor hij uiteindelijk niet gebruikt is.

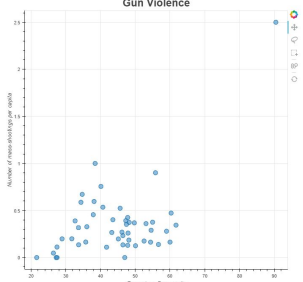


Donderdag

Technisch rapport, politiek

Uit de politieke data werd er nog een scatterplot gemaakt om de berekende correlatie nog duidelijker te visualiseren.

Verder is er opzet gemaakt voor het technisch rapport en hier vandaag verder gewerkt.



Vrijdag	Omdat het verslag van veel waarde was, is er deze dag alleen maar gewerkt aan het verslag en de opzet van de website. Dit is om te voorkomen dat het verslag op de allerlaatste dag slordig afgewerkt zou worden.
Week 4	
Maandag	<p>Er is op maandag vooral gewerkt aan het technische verslag en het design van de website.</p> <p>Ook was er een mogelijkheid om het technische rapport op te sturen naar de TA voor feedback.</p>
Dinsdag	<p>Website & technisch rapport</p> <p>Het framework voor de website is in elkaar gezet. Alle grafieken met daarbij de feitjes voor de demo zijn toen geselecteerd, zodat er de volgende dag hieraan kon worden doorgewerkt. Ook werd de feedback van Matthijs opgenomen in het technisch rapport.</p> <div><div><p>Verschillen tussen staten</p><p>Incidents per state</p></div><div><p>Incidents per capita</p></div><div><p>"Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."</p></div></div>
Woensdag	<p>De website werd afgemaakt door alle informatie als grafieken en tekst in het framework te zetten.</p> <div><div></div><div><p>Verschillen tussen staten</p><p>Incidents per state</p><p>Incidents per capita</p><p>Wie is opgenomen verschillen tussen staten?</p><p>Incidenten worden ingedeeld per regio:</p><ul style="list-style-type: none">• Midwest• South• South West• West</div><div><p>Verschillen tussen steden</p><p>Wie is opgenomen verschillen tussen steden?</p><p>De stad die het meest wapens gevonden heeft is Chicago en New York. Het aantal wapens dat gevonden is wordt weergegeven in een lijst met de stad en het aantal wapens.</p><ul style="list-style-type: none">• Chicago, New York<p>Chicago with most gun incidents 2014 - 2017</p><p>New York with most gun incidents 2014 - 2017</p></div><div><p>Wapenbezit en frequentie van incidenten</p><p>Registered gun versus incidents</p><p>Score per regio versus incidenten</p><p>Is er verband tussen de cijfers voor wapenbezit en de frequentie van incidenten?</p><ul style="list-style-type: none">• Er is een verband tussen het aantal geregistreerde wapens en het aantal incidenten, maar dit is niet een direct verband. Het aantal incidenten wordt ook beïnvloed door andere factoren.• Wanneer het gun gun incidenten meer een voor een opgenomen worden per regio.• Het is een verband, maar het aantal geregistreerde wapens is niet het enige factor dat de frequentie van incidenten beïnvloedt.</div></div> <p>De website heeft een minimalistische en schone ‘look’ gekregen, zodat de lezers/het publiek in een oogopslag de belangrijke punten herkennen. Ook worden de grafieken makkelijker te begrijpen, wanneer de pagina er niet druk uitziet.</p>
Donderdag	Het process book werd geüpdatet en afgemaakt. Verder werden de laatste puntjes op de i gezet, zoals het verschonen van de GitHub repository en het opmaken van het technisch

	rapport. Ten slotte werd de presentatie voor de demo voorbereid.
Vrijdag	De einddemo wordt (met succes) gepresenteerd.