Week 1:

We hebben gekozen voor de database die het weer van de afgelopen jaren beschrijft. Het is gelukt om de dataset naar het juiste formaat om te zetten, namelijk een .csv file. We hebben met zijn allen de data al een klein beetje bestudeert om ons een beetje te oriënteren.

Uiteindelijk hebben we met de groep toch besloten vanwege interesse en praktische redenen om voor de dataset van gun violence te kiezen. De grootste reden was dat we ons meer bij de data konden voorstellen en het leek ons daardoor makkelijker om er vragen over te verzinnen en ermee te werken.

Op het eerste gezicht leek het ons moeilijk om foute data op te sporen. Zo hebben we gekeken of er bijvoorbeeld geen absurde getallen bij het aantal doden zit en of de data kloppen. Alle getallen die we hebben gecheckt klopten en we hebben ons voorgenomen, dat we niet kloppende data later nog zouden veranderen, als we die ten minste nog gaan tegenkomen.

In de dataset stonden veel lege vakjes. Bij veel incidenten was bijvoorbeeld niet bekend welk type wapen werd gebruikt, en een hokje was dan leeg gelaten. Omdat het niet handig is om lege plekken in de data te hebben, besloten we om de lege plekken op te vullen door hier NaN in te vullen. Zo was duidelijk dat de waardes op die plekken onbekend waren. Eerst hebben we geprobeerd om dit in python te doen. We hebben de data in python geïmporteerd en de lege hokjes opgespoord. Het lukte echter niet om op de lege plekken iets in te vullen en dit naar een nieuw databestand te schrijven. Uiteindelijk besloten we om dit probleem met regex aan te pakken. We zochten hiermee weer de lege plekken op en het lukte toen wel om op de lege plekken NaN in te vullen.

Nadat de data cleaning was gedaan hebben we gediscussieerd of we ook aan data transformation moesten doen. Na hulp te vragen aan onze TA hebben we besloten dat het niet nodig was.

Het is ons nog een beetje vaag in hoeverre we de opdracht afhebben voor week 1, maar aangezien onze data al redelijk goed van zichzelf leek hebben we besloten het hierbij te laten.

Week 2:

We hebben de slides bekeken, om te zien of het van toepassing is op onze data. De slides leken vooral voorbeelden te geven over grafieken, maar het leek ons beter om zelf te beredeneren welke data visualisatie van pas kon komen. Wel hielden we de informatie van de slides in onze achterhoofd.

We hebben negen vragen verzonnen die ons interessant lijken om te beantwoorden.

1. Are there any notable differences between states/cities/years? Visualize the differences (or similarities) that you can find.

2. What are the patterns you discovered that you suspect could be interesting? Does the data contain any unusual patterns that you did not expect?
3. How does the pattern between killer/victim look between various incidents over the years. Is it mostly family/relation or do we see more reported gang violence for instance?

4 Heeft de tijd van het jaar invloed op de hoeveelheid gun violence? (Zomer, winter)
5 Kijken naar de leeftijd van killers/victims. Kijken naar geslacht van killers/victims.
6 Wat zijn de oorzaken van de uitschieters? (10 of meer doden)
7 (Kijken of bepaalde staten bepaalde gun_types gebruiken)
8 (Verhouding killed/injured bij man of vrouw/jong of oud, als man 10 shiet doodt ie meer dan als een vrouw 10 keer schiet)
9 Hoe zal gun violence in de VS zich in de komende jaren ontwikkelen?

Voor je vraag 2 echt kan beantwoorden, moet je natuurlijk eerst data onderzoeken. Wij hebben dus eerst data onderzocht waarvan we dachten dat het interessant kon zijn. Dit kwam uit op het vergelijken van verschillende gebieden in Amerika.

We hebben verzonnen dat iedereen zich op bepaalde vragen focust en dat wanneer er vragen zijn we er met zijn allen naar gaan kijken. Als iemand zijn eigen vraag af heeft kiest hij een nieuwe vraag totdat hopelijk uiteindelijk alle vragen op zijn.

Voor de overzichtelijkheid van het logboek kiezen we ervoor dat elke vraag zijn eigen kleine logboek krijgt. Daarbuiten schrijven we nog steeds per week wat er gebeurde buiten de vragen om (dus bv wat we met de hele groep gedaan hebben). De kleine beschrijving van hoe we de vragen hebben beantwoord staan helemaal onderaan.

Week 3:

We zijn vooral individueel verder gegaan met het verwerken van de vragen. Aangezien we redelijk vaak samen bij elkaar zaten op school was het voor ons niet nodig om github veel te gebruiken. Rond deze tijd hebben we dus vrij weinig bestanden gepusht, terwijl we wel redelijk hard aan het werk waren.

Aangezien het grootste gedeelte van deze week m zat in het verdiepen en verwerken van de vragen, zijn er weinig gezamenlijke beslissingen gemaakt. Wel hebben we met zijn allen een beetje bedacht hoe de laatste week moest verlopen,aangezien het technisch rapport en de website nog gemaakt moesten worden en dit waarschijnlijk best veel werk in beslag zou nemen.

Week 4:

Wij schrijven allemaal stukjes voor het technisch rapport. Iedereen schrijft over de vraag waarin hij zelf zich in verdiept heeft. We hebben ook twee mensen de taak gegeven om zich te oriënteren in de werkingen van de website.

We zijn de week vooral allemaal naar elkaars werk aan het kijken en zijn aan het verzinnen hoe alles in een goed en duidelijk verhaal kan passen.

Vraag 1:

Om een algemeen beeld te krijgen van de mate van gunviolence hebben we gekeken naar het aantal doden en gewonden per staat en per jaar.

We hebben een apart csv bestand aangemaakt en daar de staten ingezet met bijbehorende waardes. District of Columbia is echter geen staat, hoewel deze wel bij de staten stond in het originele bestand, en omdat we besloten hadden de data in een kaart van de Verenigde Staten weer te geven, hebben we het genegeerd.

Uit de resultaten kwamen bij een aantal Staten dezelfde waardes. Dit hebben wij zo goed mogelijk proberen te verhelpen door naar het inwoneraantal te kijken van deze staten en daarop gebaseerd een ranglijst te maken.

Het verschil tussen het aantal doden en gewonden per jaar is weergegeven in een staafdiagram om zo makkelijker patronen te herkennen.

In de onderzoeksvraag 1 zou er in eerste instantie ook gekeken worden naar de verschillen tussen de steden in de VS. Wij hebben ervoor gekozen dit uiteindelijk niet te doen, omdat de staten op zich al een goed genoeg beeld gaven van het wapengebruik per gebied.

Vraag 2:

Vanwege de amerikaanse ruzies tussen oostkust en westkust en de vooroordelen van staten als Texas, hebben we het aantal incidenten per verschillende gebieden bekeken.

Het onderverdelen in lengte en breedte graden van amerika ging nog wel, maar blijkbaar er zitten ook punten tussen die foutief zijn. Er is een punt dat bijvoorbeeld midden in Azië lag. We hebben besloten die punten gewoon te negeren.

Er komen nogal rare resultaten uit, maar aangezien we geen fouten in de simpele duidelijke code kunnen vinden denken we dat het of aan de data moet liggen of dat de waarheid onverwachts is.

Aangezien uit de eerste methode geen goede resultaten komen hebben we ervoor gekozen een nieuwe methode te kiezen. Deze methode is gebaseerd op het gebieden indelen in staten en het delen door het aantal inwoners. De resultaten die we kregen werden hierdoor al een stuk logischer.

We hebben besloten de data in staafgrafieken weer te geven.

Vraag 3:

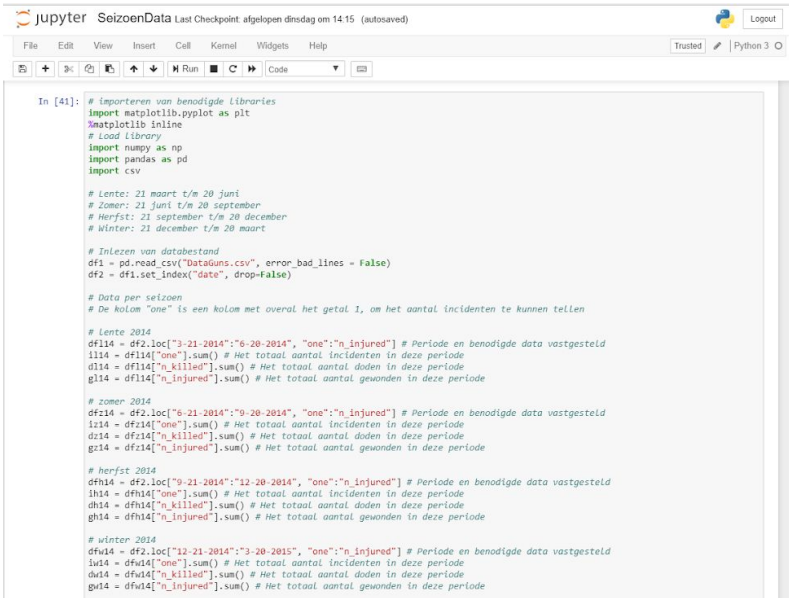
Bij vraag 3 hebben we zes relaties tussen schutter en slachtoffer bekeken waarvan we dachten dat ze belangrijk waren. Sommige daarvan leken onszelf belangrijk om te testen, zoals gang violence, aangezien ze vaak in het nieuws komen. Andere hebben we gekozen omdat ze heel vaak voorkwamen zoals family.

Uiteindelijk hebben we via de slides gekozen voor een ‘100% stacked bar chart’. Dit is omdat de data van 2018 niet compleet was, aangezien het jaar nog niet volledig is afgelopen. Maar omdat de informatie van dat jaar wel gewoon juist is, kan je als je alleen naar de verhoudingen van dat jaar kijkt, er nog veel informatie uit halen.

We hadden eigenlijk een doel om te voorspellen hoe de verhoudingen in de toekomst zouden verlopen. Dit wilden we doen met ‘multi variate linear regression’. Het is uiteindelijk niet gelukt om dat voor elkaar te krijgen.

Vraag 4:

Om naar de invloed van de tijd van het jaar op gun violence te kijken, moest het jaar in stukken worden verdeeld. Een handige verdeling leek ons de seizoenen, deze staan in verband met het weer en wij verwachten dat dit de belangrijkste factor is als het gaat om de invloed op gun violence. Omdat de data van 2013 onvolledig was is er gekozen om deze data buiten beschouwing te laten. Dit zou anders voor rare verhoudingen kunnen zorgen tussen verschillende seizoenen en jaren. Omdat de eerste onderzochte winter ook een deel van 2013 bestrijkt hebben we ervoor gekozen om pas naar de data vanaf de lente in 2014 te kijken. Dit kwam ook goed uit omdat de data van 2018 tot en met 31 maart liep (deze was verder wel volledig). Zo is de data van vier keer vier seizoenen bekeken, wat resulteerde in 16 variabelen met daarin alle gun violence incidenten in die periode. (De seizoenen zijn als volgt ingedeeld: lente: 21 maart t/m 20 juni, zomer: 21 juni t/m 20 september, herfst: 21 september t/m 20 december, winter: 21 december t/m 20 maart.)



Om het aantal incidenten in een periode te tellen, kon er geen simpele functie als length worden gebruikt omdat de variabelen grote tabellen met veel rijen maar ook kolommen waren. Hierom besloten we om voor deze dataset een kolom te maken met overal het getal 1 erin (de kolom “one”). Op deze manier kon de sum functie worden gebruikt om het aantal incidenten te tellen. Het aantal doden en gewonden kon ook met de sum functie worden berekend. Al deze waardes konden vervolgens in een staafdiagram worden gezet. Omdat er hier niet duidelijk uit bleek wat de verhouding was tussen verschillende seizoenen besloten we om ook cirkeldiagrammen te maken van het aantal incidenten, doden en gewonden en nog een cirkeldiagram met alles bij elkaar voor de volledigheid.

Vraag 7:

We begonnen met de vraag of er een patroon te zien was tussen de soort wapens en de staten. Hieraan zijn we begonnen door eerst alle wapens die voorkwamen in de dataset op een rijtje te zetten. Om dit te doen hebben we eerst alle data moeten filteren op Unknown values en vakjes met ‘NaN’. Omdat elk wapen in een bepaalde format stond (0::Wapen1||1::Wapen2) moesten we een code verzinnen die deze format omzeilde. Dit hebben we uiteindelijk met regex opgelost.
Nadat we erachter zijn gekomen hoe we de wapens uit de format konden halen, konden we aan de slag met het tellen van alle wapens per staat.
Op de eerste blik was de data waardeloos, aangezien bij elke staat de handgun, shotgun, rifle en 9mm ruim bovenaan stonden. Uiteindelijk hebben wij ervoor gekozen deze te negeren, omdat het algemene namen waren voor veel andere wapens. Hierna hebben we een tabel gemaakt die van elke staat de top 3 gebruikte wapens weer gaf. Omdat deze tabel zeer onoverzichtelijk was hebben we gekozen voor een staafdiagram dat het aantal keer dat een wapen het meest gebruikt werd in een staat aangaf. De vraag hebben we tegelijk ook veranderd omdat er geen duidelijk patroon te herkennen was tussen staat en geweer. De vraag werd: Wat zijn de meest populaire wapens in Amerika?

Vraag 9:

Vraag 9 ging heel soepel. We hebben de automatische linear regression van sklearn gebruikt. Ook was dit de laatste vraag waar we aan werkte, vandaar dat alles dus vrij gemakkelijk ging.