Sumário Executivo

Os EUA são um dos poucos países em que o direito de portar armas é constitucionalmente protegido. Uma consequência disso é que o nível de violência armada é muito alto em comparação com outros países. Depois de analisar o conjunto de dados, que foi baixado de gunviolence.org, minhas principais descobertas foram:

* A tendência é, infelizmente, para cima. O número de incidentes aumentou ano-a-ano de 2013 para 2017. O número de incidentes no Q1 2018 é menor que esse número no Q1 2017, que pode ser visto como um sinal cautelosamente positivo.
* Em média, o número de incidentes é menor no outono/inverno do que na primavera/verão. Fins de semana são mais perigosos que os dias de semana, mas isso era óbvio para mim. As datas mais perigosas são 1 de Janeiro (dia de ano novo), e 4 de julho (dia da independência, com muitos incidentes também acontecendo no dia seguinte, 5 de julho).
* Eu enriquecia os dados com números populacionais do departamento de Censos dos EUA. Depois de ter feito isso, o Alasca saiu como o (para mim surpreendente) estado com o número relativamente maior de incidentes. Além de Alaska, Louisiana e Delaware são também Estados com números médios elevados do incident. O estado mais seguro sobre o número relativo de incidentes é o Havaí. No entanto, o Alasca tem uma relação relativamente baixa de vítimas (número de pessoas mortas + feridos) por incidente. Isso parece ser causado por uma alta percentagem de "não-tiroteio incidentes" (presumo que isso significa principalmente incidentes onde as pessoas só ameaçaram atirar). Ao classificar os Estados sobre o número de vítimas em relação ao tamanho da população, Louisiana é o estado mais perigoso. Nesta estatística, Illinois salta para o segundo lugar, que é devido ao Estado ter a maior proporção de vítimas por incidente.
* Das cidades grandes com pelo menos 600.000 habitantes, Baltimore, Washington DC, e Chicago têm o número relativo o mais elevado dos incidentes. As cidades grandes são geralmente mais perigosas do que o lado do país a respeito da violência do injetor, porque as taxas do incident são (muito) mais elevadas.
* O incidente mais grave de longe foi o tiroteio na Strip em Las Vegas (a partir do Hotel Mandalay Bay), com um total de 470 pessoas que foram mortas ou feridas.
* As características do incidente apresentam uma riqueza de informações. Ao todo 110 categorias foram usadas para descrever os incidentes. Em alto nível, isso pode explicar, por exemplo, a relação baixa das vítimas/incidentes de Alaska. Entretanto, categorias como Terrorismo, a gangue, e a participação da droga foram registradas igualmente.
* Eu incluí um mapa que resume os incidentes relacionados às drogas. Acontece que Baltimore tem muito mais incidentes relacionados às drogas, quando comparado com outra cidade de mesmo tamanho (citação Internet: em uma cidade de 645.000, o departamento de Baltimore de Saúde estima que há 60.000 viciados em drogas, sendo 48.000 deles viciados em heroína).
* Em relação aos acidentes com gangues, Chicago é a cidade que tem de longe o maior número de incidentes.
* 30 incidentes foram rotulados como "envolvimento do terrorismo", e eu também os apresentei em um mapa.

Um dos meus objetivos desta análise foi aprender a fazer mapas interativos em R, e os meus primeiros mapas interativos com Leaflet foram incluídos nesta versão. Além disso, também converti alguns objetos ggplotspara plotly , que permite a exibição interativa de rótulos com informações exatas.

Introdução

Gun Violence Archive (GVA) é uma organização sem fins lucrativos criada em 2013 para fornecer acesso público on-line gratuito à informações sobre violência relacionadas às armas nos Estados Unidos. A GVA afirma que a América é um país excepcional quando se trata de armas. É um dos poucos países em que o direito de portar armas é constitucionalmente protegido. Mas a relação da América com as armas é única em outra forma crucial: entre as nações desenvolvidas, os EUA são de longe os mais homicidas. Por exemplo, a América tem seis vezes mais homicídios de armas de fogo do que o Canadá, e quase 16 vezes mais que a Alemanha.

A data nesse conjunto de dados foi baixada do [gunviolencearchive.org](http://www.gunviolencearchive.org/). O arquivo CSV contém dados para todos os incidentes registrados de violência armada nos EUA entre janeiro de 2013 e março de 2018. Os dados consistem de 240K incidentes, com informações detalhadas sobre cada incidente.

Lendo e Explorando Dados

# Carregando bibliotecas necessárias e lendo os dados em R

Carregando pacotes R usados além da base R

Nos casos em que precisei apenas de uma função específica de um pacote, decidi não carregar o pacote inteiro, apenas carregar a função que preciso usando a notação package :: function ().

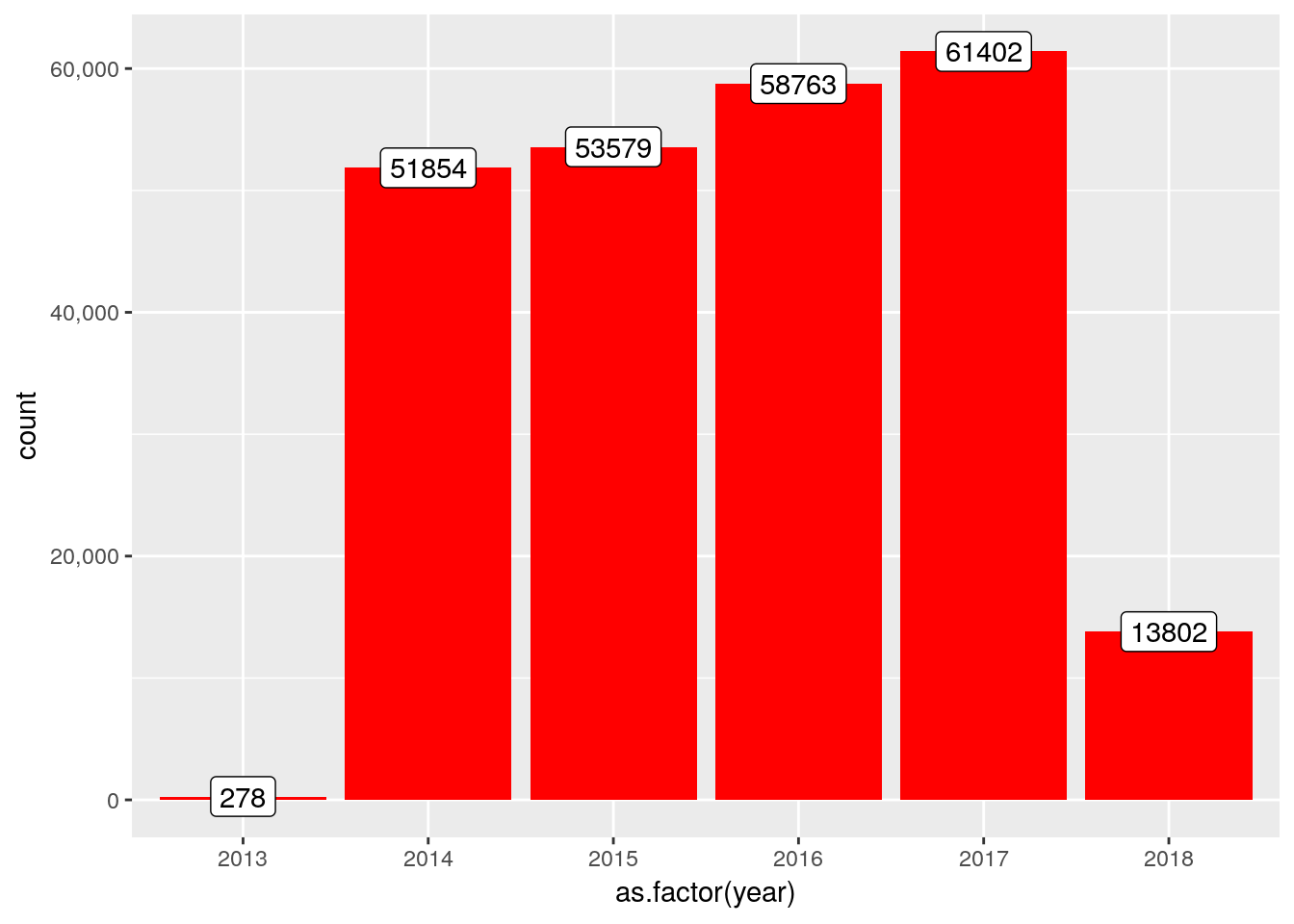
Foi adicionado manualmente o incidente de Las Vegas que faltava. O tópico *[Las Vegas incident missing](https://www.kaggle.com/jameslko/gun-violence-data/discussion/55307)* na lista de discussão explica o porquê o incidente em questão foi retirado.

Tamanho e Estrutura dos dados

Muitas variáveis precisam ser reformatadas antes de poderem ser usadas. Eu escolhi fazer isso por variável assim que precisar da variável para uma exploração particular.

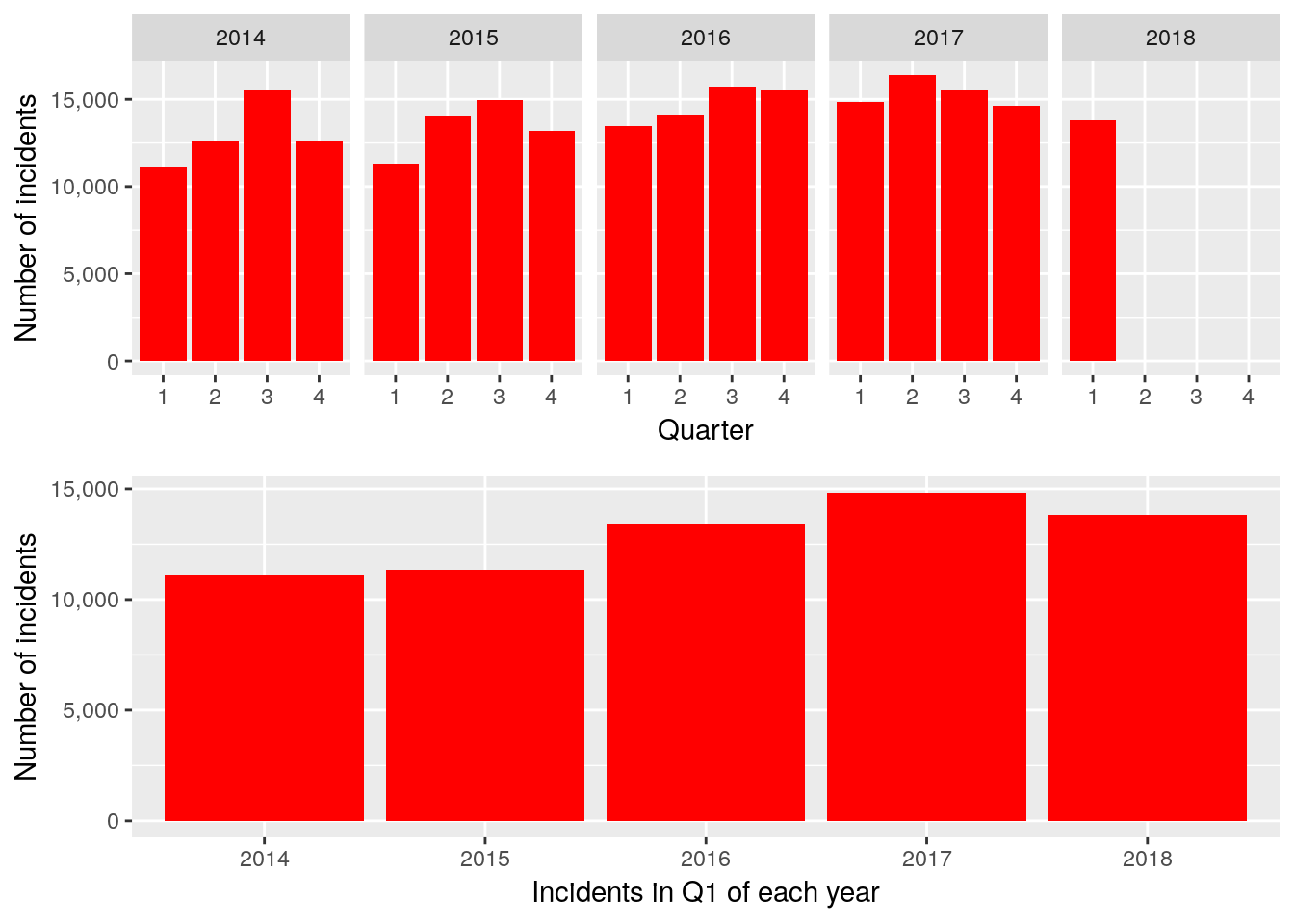
# Por Ano

Parece que a maioria dos incidentes em 2013 não foi registrada. Além disso, 2018 não é um ano cheio de dados, pois o último incidente registrado foi em 31 de maio de 2018. Como isso significa exatamente um quarto, estou interessado se a tendência ainda é para cima ao olhar apenas para os Q1s de 2014-2018. Vou analisar isso na próxima seção.



# Por trimestre; o que o Q1 2018 nos diz?

Parece haver algum tipo de "sazonalidade", com Q1 e Q4 geralmente apresentando números menores de incidentes do que Q2 e Q3. O segundo gráfico mostra que o primeiro trimestre de 2018 tem menos incidentes do que o primeiro trimestre de 2017. Isso pode ser visto como um sinal cautelosamente positivo. No entanto, deve-se perceber que esse número ainda é muito alto quando comparado a outros países (relativamente).

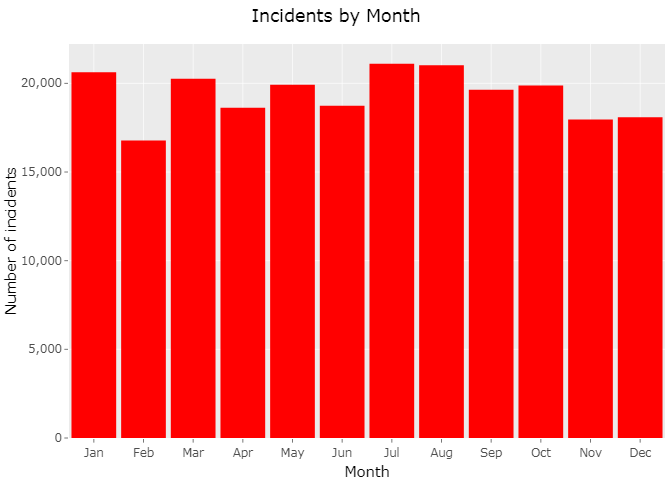


# Por Mês

A análise dos trimestres mostra que mais incidentes ocorrem nos períodos mais quentes de primavera e verão. Isso parece valer a pena mergulhar um pouco mais fundo. Para comparar os meses, vou excluir 2018, já que apenas os três primeiros meses foram registrados.

Abaixo, o efeito mais visível da "sazonalidade" me parece que os meses mais frios parecem ter menos incidentes. Novembro, dezembro e fevereiro são os 3 meses com o menor número de incidentes (fevereiro também tem apenas 28 dias de curso). A exceção parece ser janeiro, o que vale a pena investigar mais tarde. Minha primeira ideia é que possivelmente incidentes na véspera de ano novo contribuam para que Janeiro tenha um grande número de incidentes.

O outro pico é o período de julho / agosto. Provavelmente porque muitas pessoas vão sair de férias neste período.



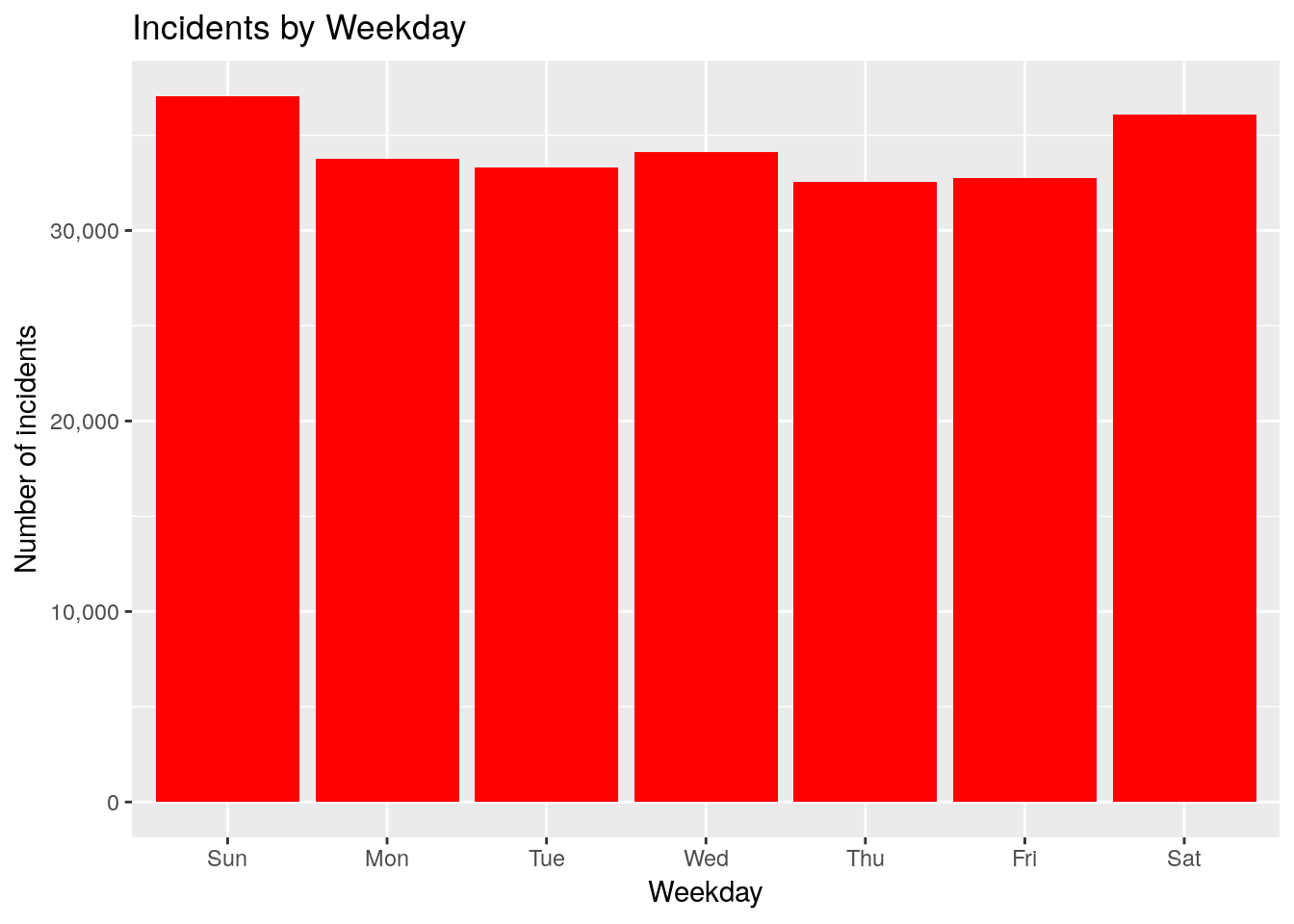
# Datas com mais incidentes

Como eu mencionei a passagem de ano pode ser uma possível explicação dos altos números de janeiro, enquanto os meses mais frios como novembro a fevereiro parecem ter menos incidentes, queria verificar quais são as datas com a maioria dos incidentes.

Os números acima são realmente os totais de 4 datas, pois são agregados em 4 anos (por exemplo: 1-1-2014, 1-1-2015, 1-1-2016, 1-1-2017). Com a média sendo 618 (o número total de incidentes em 2014-2017 dividido por 365 dias corridos; 225598/365), não há muitas datas que realmente se destacam. A maioria das datas neste top 10 parecem "comuns" em julho / agosto. No entanto, em 1º de janeiro, de fato, explica parcialmente os números mais altos de incidentes em janeiro. Além disso, o dia da independência (4 de julho) também é perigoso quando se trata de incidentes com armas de fogo. Eu suponho que o número alto em 5 de julho se deve ao fato de as pessoas continuarem a comemorar depois da meia-noite.

# Por dia da semana

Fins de semana (sábado e domingo) são mais perigosos do que os dias úteis. Meu melhor palpite é que isso é devido a maioria das pessoas não ter que trabalhar nestes dias, e muito provavelmente devido à violência relacionada à vida noturna. A violência que acontece na sexta-feira à noite provavelmente acontecerá depois da meia-noite e, portanto, será registrada como evento aos sábados.

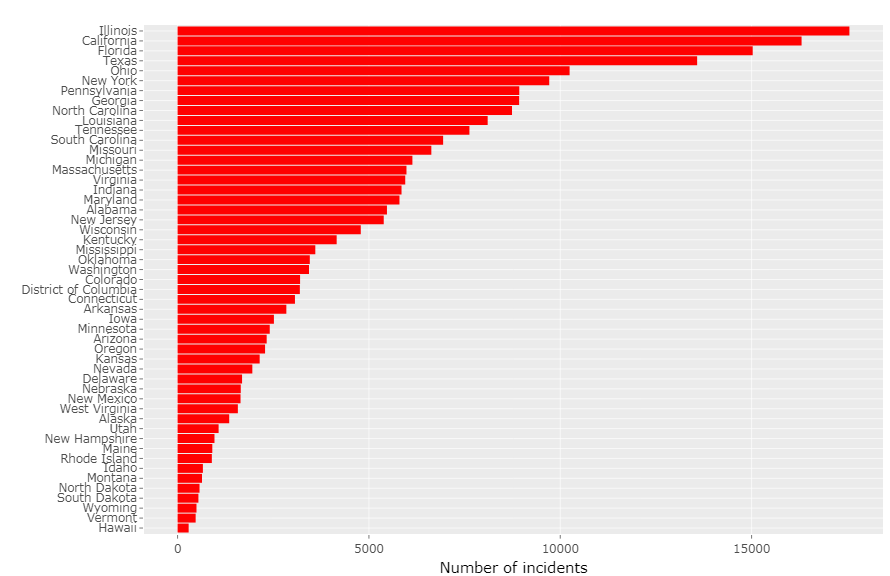


# Comparando o número de incidentes e vítimas por localização

### Incidentes por estado

Primeiro de tudo, eu tenho que converter o estado em uma variável de fator. Como eu também analisarei por cidade ou por estado mais tarde, eu vou converter os dois ao mesmo tempo.

Abaixo, estou exibindo os números absolutos de incidentes por estado. No entanto, esses números devem estar relacionados ao número de habitantes para obter uma boa visão do número relativo de incidentes. Por exemplo, a Califórnia é um estado com uma população muito grande. Portanto, não é surpresa ver a Califórnia como o segundo colocado em números absolutos.



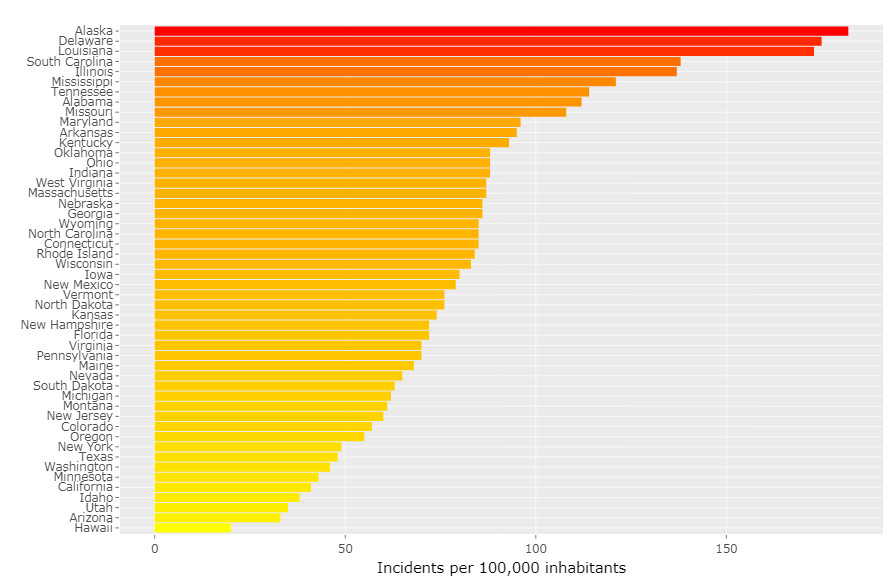
### Incidentes relativos ao tamanho da população do Estado

Para poder relacionar os números acima com os tamanhos populacionais, tive que procurar uma boa fonte. Eu encontrei o que eu precisava no site do US Census Bureau. A leitura direta dos dados em um dataframe não funcionou no kernel do Kaggle. Portanto, eu fiz o upload para um conjunto de dados privado (o link para ele é definido no início do documento. path2 <- “../input/erik-gun-extra/”).

Abaixo, estou criando um pequeno dataframe de "estados", que eventualmente contém o número de incidentes em relação à população de cada estado (por 100.000 habitantes).

Na figura abaixo, você pode ver que os dados enriquecidos, relacionados à população de cada estado, apresentam um quadro muito diferente. Como o número de incidentes está relacionado ao tamanho da população, esses números agora representam níveis de perigo de arma "reais". Para mostrar isso visualmente, usei códigos de cores. Vermelho significa um alto nível de perigo em termos de números relativos de incidentes, e amarelo significa que um estado é relativamente seguro.

Agora, o Alasca, a Louisiana e o Delaware estão mostrando os números de incidentes relativos mais altos. O Havaí parece ser o estado mais seguro para se viver (em relação a incidentes relacionados a armas, terremotos e erupções vulcânicas são uma história diferente :-)), e o grande estado da Califórnia cai do segundo estado em termos de incidentes absolutos para uma posição baixa quando corrigido (relativizado) à sua grande população.



Nota: na verdade, o Distrito de Columbia saiu como o estado com a maior taxa de incidentes. Com toda a honestidade, o estado "Distrito de Columbia (DoC)" não tocou nada para mim. No entanto, uma pesquisa rápida na Internet me disse que não é realmente um estado, mas na verdade um “distrito federal” não relacionado a nenhum estado e aproximadamente igual à cidade de Washington DC (veja também Huffington Post, Tripsavvy). Como esses incidentes têm que ser atribuídos de qualquer maneira, eu entendo que o DoC é rotulado como um estado nos dados. No entanto, como Washington DC já é tratada como uma cidade na seção 4.2.3.1, decidi retirá-la do gráfico acima.

O estado do Alasca com relativamente mais incidentes foi uma surpresa para mim! Claro, eu queria descobrir de onde esse alto nível de violência armada no Alasca poderia vir. O link a seguir explica por que os dados deste conjunto de dados não estão mentindo: é por isso que o problema com o canhão do Alasca é tão ruim. Em resumo, o Alasca tem leis de armas muito liberais, e índice de posses de armas muito altas como consequência (60% dos lares), e também um nível muito alto de suicídios (80% das mortes por arma de fogo são auto-infligidas).

Para ter uma ideia de como essas figuras relativas são compostas, estou exibindo os números subjacentes de alguns estados abaixo. Como você pode ver, o DoC e o Alasca têm poucos habitantes, mas um número relativamente alto de incidentes. O Havaí tem poucos números de incidentes, e o grande estado da Califórnia apresenta um número relativamente baixo de incidentes ao levar em conta a população.

### A alta taxa de incidentes no Alasca é estatisticamente significativa?

Se número mais alto do Alasca fosse significativo, devido ao tamanho relativamente pequeno da população. Para descobrir isso, adicionei um teste z (ou teste de hipótese) de 2 proporções.

O p-value deste teste é muito pequeno, portanto podemos concluir que a taxa de incidentes no Alasca é significativamente maior que a taxa do resto dos EUA. As estimativas da amostra devem ser lidas como: prop1 é a estimativa do Alasca. Se multiplicarmos esse número por 100.000, estaremos obtendo os 182 incidentes por 100.000 habitantes. O resto dos EUA tem uma média de 73 incidentes por 100.000 (prop 2 \* 100.000).

O intervalo de confiança deve ser lido da seguinte forma:

* a estimativa pontual da diferença é 182-73 = 109
* com 95% de confiança, a diferença real deve ser entre 99 e 118
* o fato de que esse intervalo de confiança não inclui 0 também significa que a diferença é significativa.

Como as margens estatísticas de erro são razoavelmente pequenas, cerca de mais ou menos 10 para um estado pequeno como o Alasca, não vou incluir mais nenhum teste estatístico nesta análise.

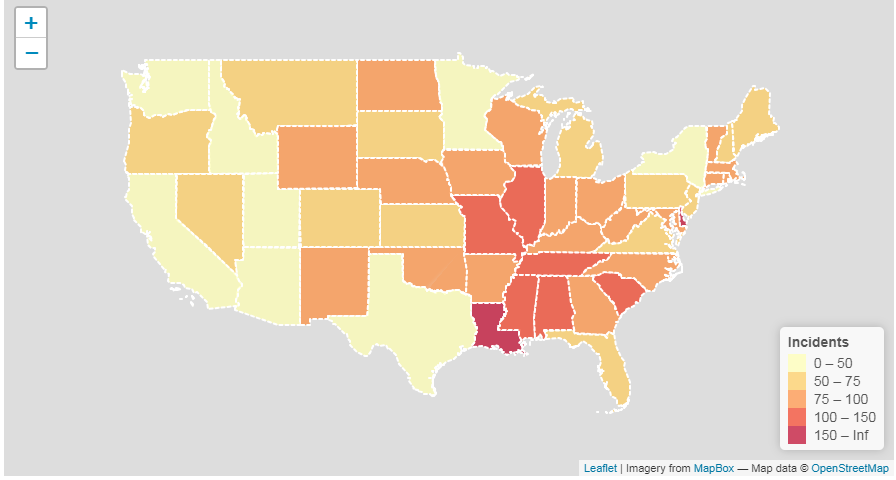
### Um mapa interativo de incidentes por estado com Leaflet

Como mencionado no Sumário Executivo, um dos meus objetivos ao adotar este conjunto de dados foi aprender a criar mapas interativos. Consegui fazê-lo e estou planejando fazer muito mais mapas interativos no futuro. Eu fiz o download dos dados espaciais da agência do Censo dos EUA e carreguei os arquivos descompactados no conjunto de dados privado (path3 <- “../input/erik-gun-extra”). Abaixo, estou usando o pacote rgdal para ler esses dados espaciais.

Esse dataframe de polígonos espaciais também contém elementos de dados, como você pode ver abaixo.

Com essa estrutura de dados, uma junção simples (unindo a coluna Per100000) com dados adicionais não é possível. No entanto, descobri que uma solução fácil (por enquanto) é extrair elementos de dados do objeto, junte-se a esse pequeno dataframe com as informações que desejo exibir no mapa (os incidentes por 100.000 habitantes) e incluí esta coluna no objeto 'estados'.

Agora é hora de criar o mapa. Sinta-se à vontade para passar o mouse sobre os estados (isso mostra um pop-up com o nome do estado e o número exato de incidentes por 100.000 habitantes) e faça zoom in e out. Ao diminuir o zoom, você poderá ver o Alasca e o Havaí. Ao aumentar o zoom, você também pode encontrar o Distrito de Columbia (entre Maryland e Virgínia).

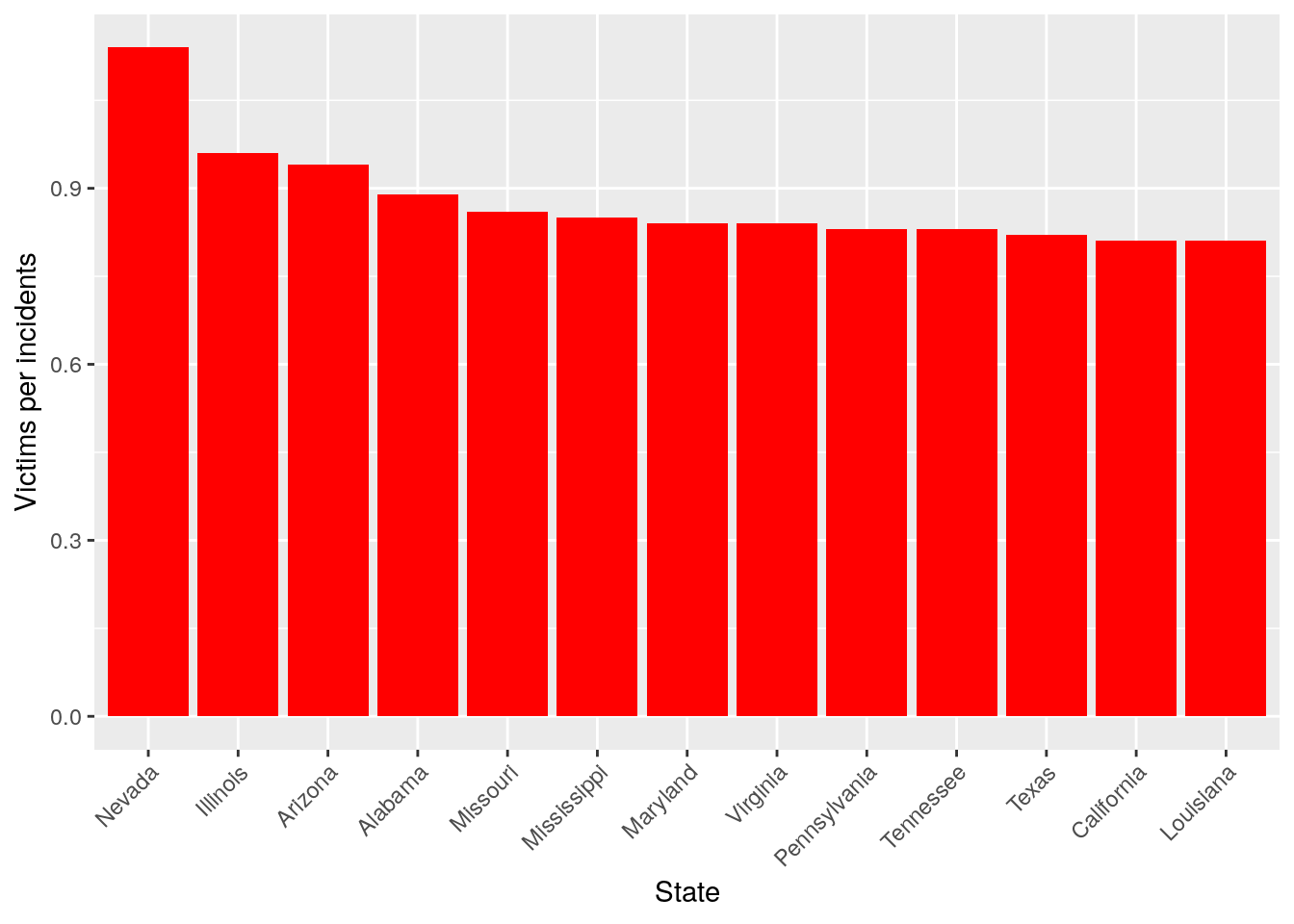


# Vítimas por Estado

Primeiro de tudo, vou criar uma nova variável que adiciona o número de pessoas mortas e feridas para cada incidente.

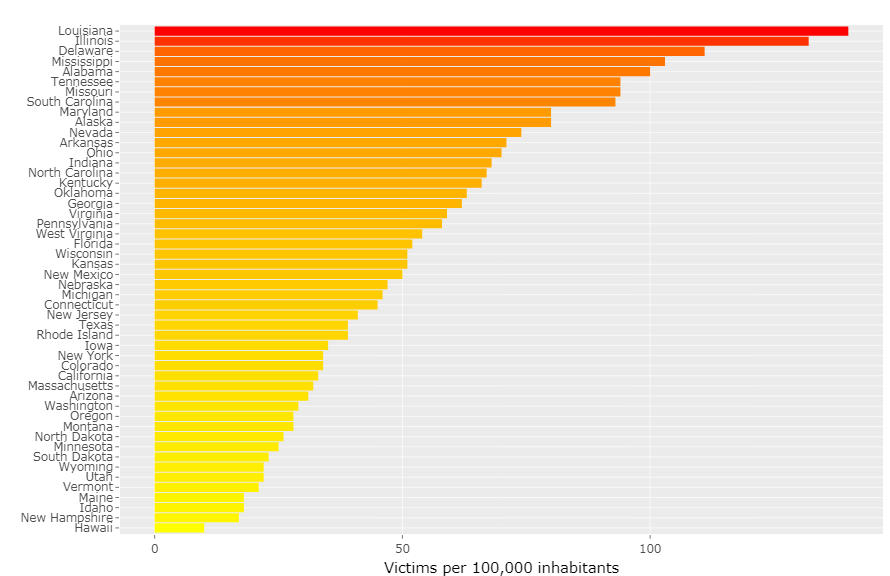
### Gravidade dos Incidentes

Como você pode ver na tabela acima, o Alasca tem um número relativamente baixo de vítimas por incidente (0,44). No entanto, como mostrado no gráfico, a Louisiana também está em alta em relação a essa relação.



### Vítimas em relação ao tamanho da população do estado

No gráfico abaixo, que novamente usa os códigos de cores para mostrar os níveis de perigo, você pode ver que a Louisiana é o estado mais perigoso quando se trata do número de vítimas em relação ao tamanho da população. Isso não é surpresa, já que o estado tem um alto número relativo de incidentes e também é alto em relação à proporção de vítimas por incidente. O Alasca cai para o 9º lugar, o que se deve ao número relativamente baixo de vítimas por incidente.



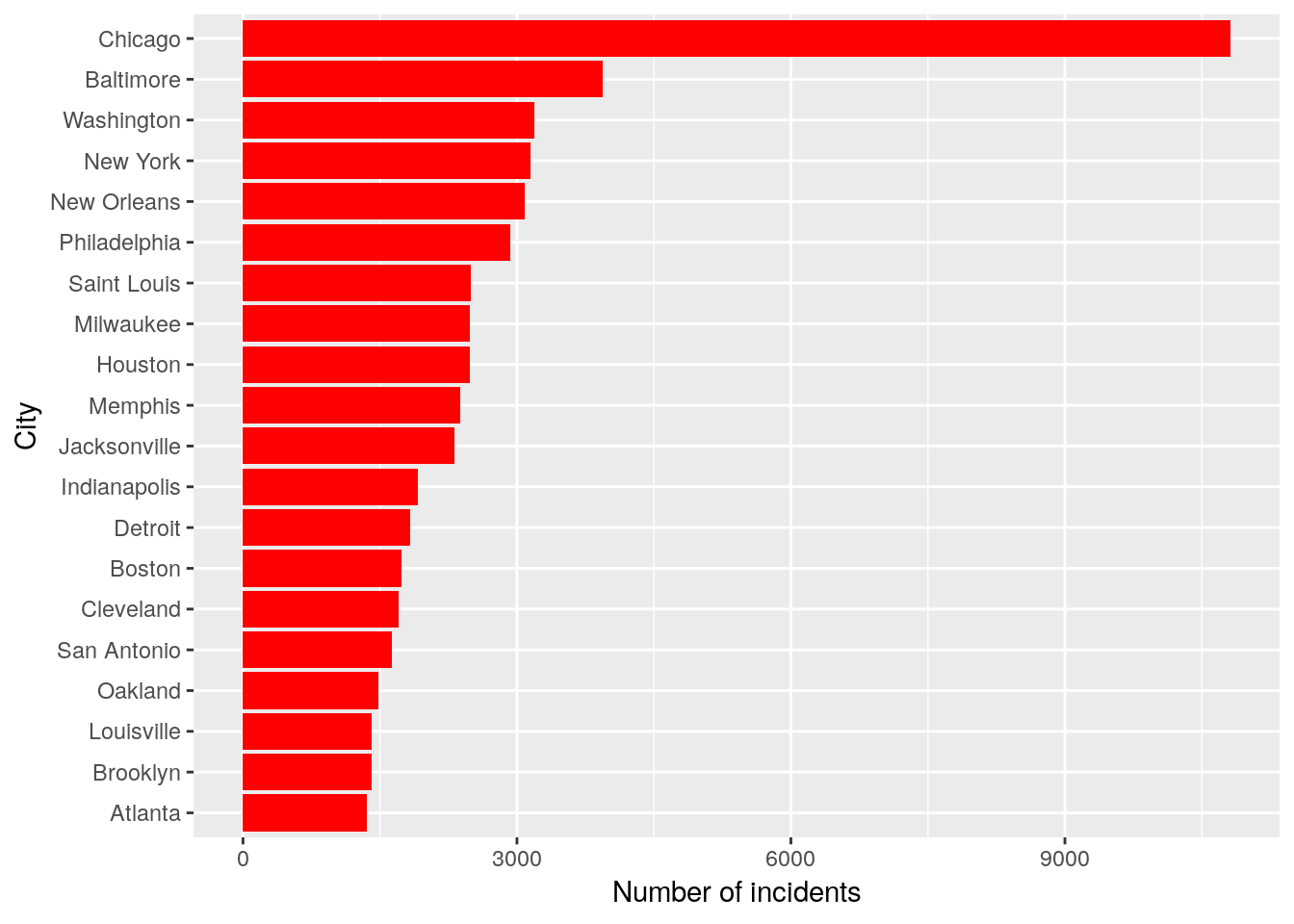
# 

# Incidentes por cidade

Parece provável que muita violência ocorra nas grandes cidades dos estados. Por exemplo, eu esperava ver muitos incidentes nas grandes cidades de Illinois (Chicago) e Louisiana (Nova Orleans). É improvável que a maior cidade do Alasca (Anchorage) apareça na lista de cidades com o maior número absoluto de incidentes, já que a cidade é relativamente pequena.

A cidade de Nova York não parecia estar registrada como um todo no conjunto de dados. Acontece que os bairros são registrados separadamente.

Antes de continuar, estou adicionando Nova York como cidade.

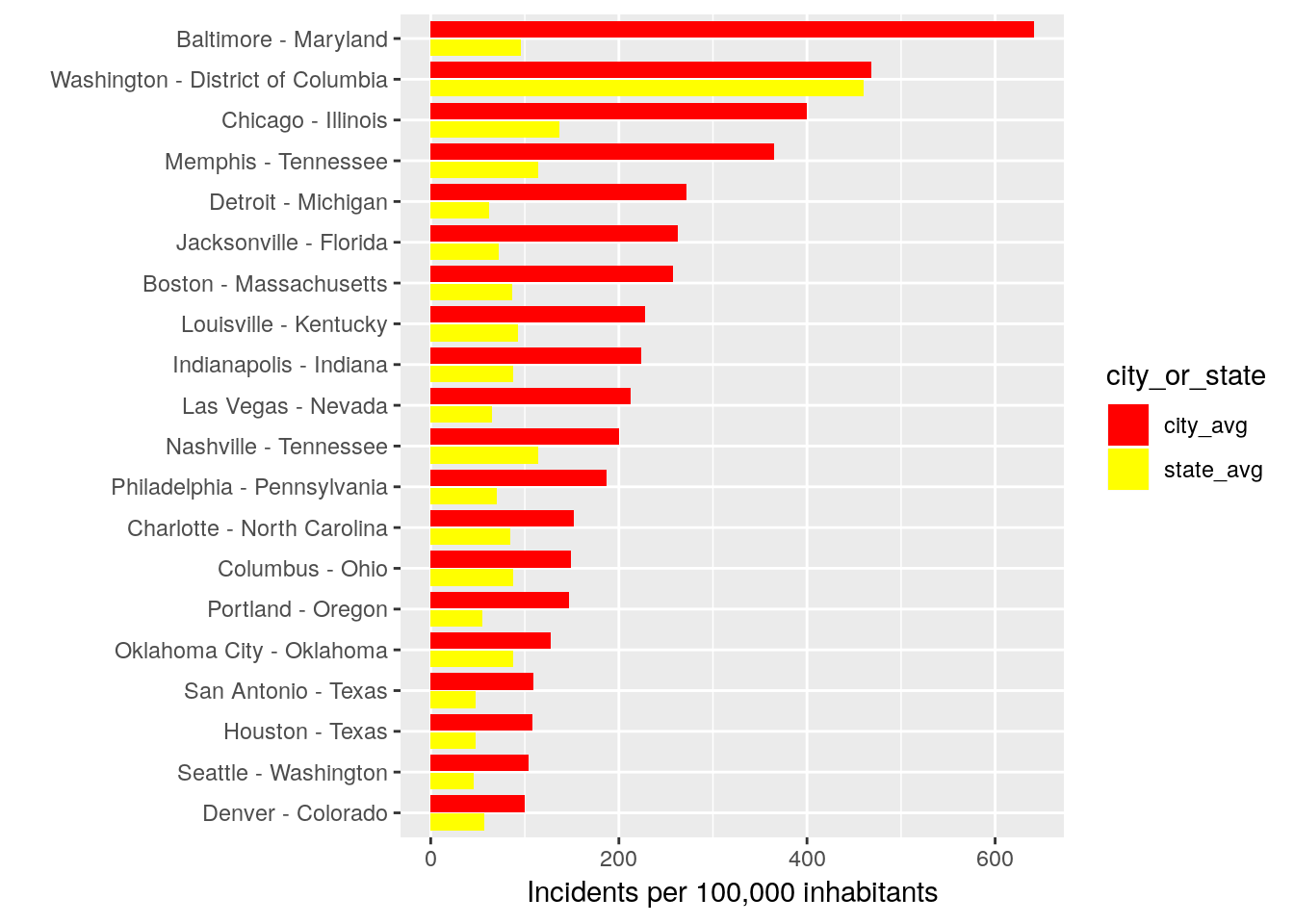


Embora eu esperasse que Chicago tivesse muitos incidentes, ainda estou surpreso em ver que ele detém o "primeiro lugar" por uma enorme margem! Na próxima seção, relacionarei esses números absolutos à população por cidade.

### Incidentes relativos ao tamanho da população da cidade

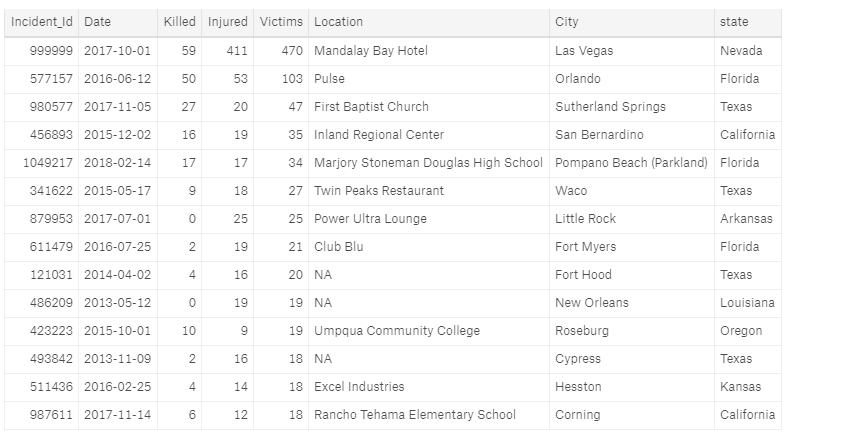
Como esse é um arquivo grande, decidi me concentrar apenas nas cidades maiores. Selecionar todas as cidades com pelo menos 600.000 habitantes inclui apenas Baltimore; o número 2 em números absolutos de incidentes.

Como você pode ver, as médias das cidades com números altos de incidentes são muito mais altas do que as médias dos respectivos estados. Na parte inferior desta lista (não exibida), algumas cidades têm números menores que o estado, mas geralmente as cidades são mais perigosas do que a média do país.

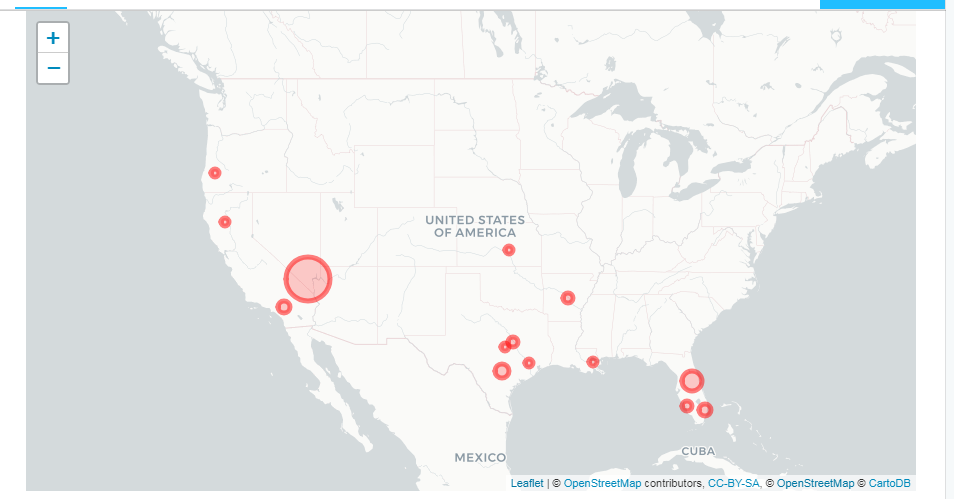


### Incidentes com maior número de vítimas

Abaixo, estou exibindo os 13 incidentes com o maior número de vítimas. Embora uma única vítima seja uma em excesso, o tiroteio em Las Vegas foi de longe o pior incidente com mais de 470 vítimas.



### Um mapa interativo dos incidentes com maior número de vítimas

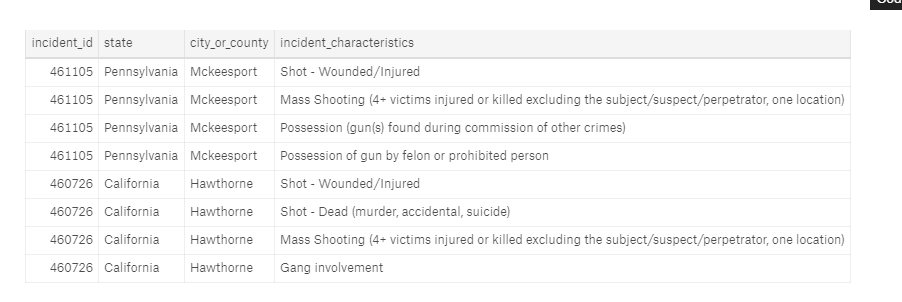


Analisando as Características do Incidente

Como você pode ver abaixo, o campo "incident\_characteristics" realmente contém muita informação. Parece que a pessoa que inseriu as informações teve a opção de selecionar várias categorias pré-descritas. Estes são separados por "||" na coluna incident\_characteristics. Os dois primeiros incidentes são ambos descritos por quatro categorias, mas isso varia entre os dados.

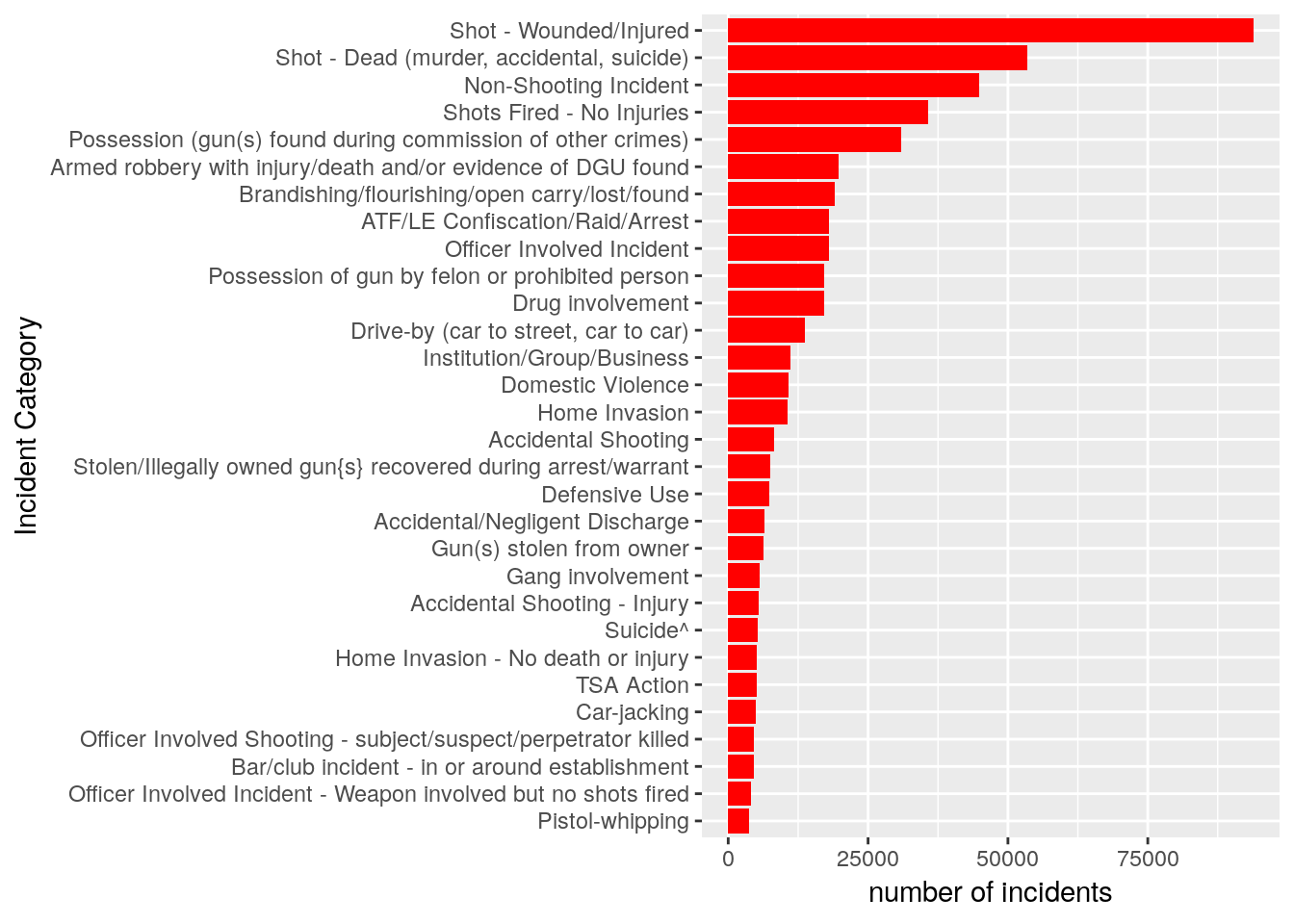
A função cSplit da forma splitstacks do pacote funciona perfeitamente para dividir essa coluna e armazenar todas as características como observações separadas em um novo dataframe. Além do incident\_id, também estou mantendo o estado e a cidade, pois também quero ver os detalhes específicos, por exemplo, Alaska e Baltimore.

Abaixo, você pode ver que os dois primeiros incidentes estão divididos em oito observações.



# Categorias de incidentes nos EUA

Como você pode ver, existem muitas categorias diferentes. Estas não necessariamente envolvem tiros sendo disparados. Por exemplo, suponho que "Incidente Sem Tiro" significa que as pessoas apenas ameaçaram atirar, ou possivelmente usaram sua arma como forma de impressionar.

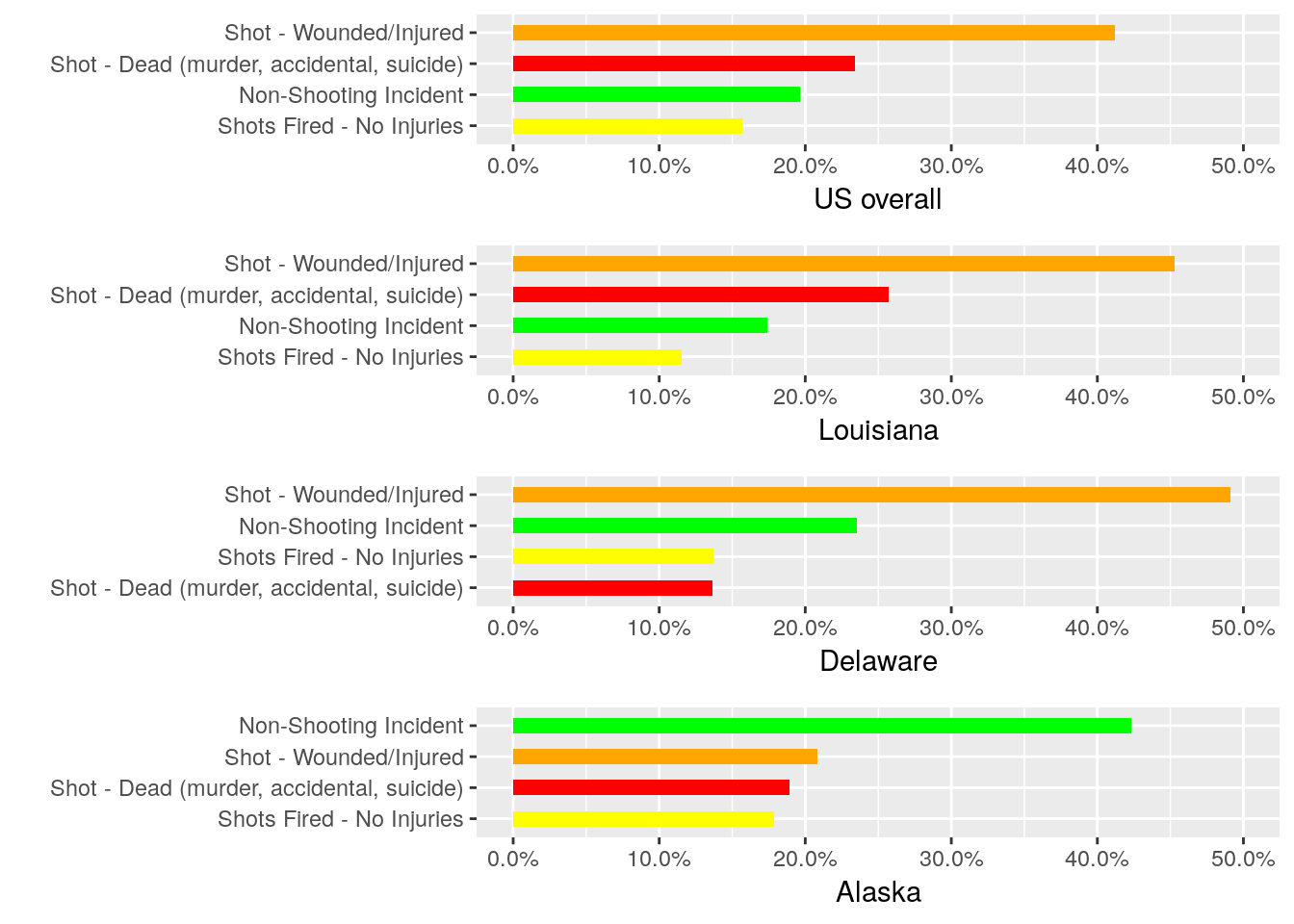


Como as primeiras 4 categorias parecem categorias globais, exibirei números nessas 4 categorias separadamente no restante desta seção.

Na minha opinião, isso está perto o suficiente para 100%. Eu estou supondo que a distribuição sobre essas 4 categorias para os restantes 5% dos incidentes é aproximadamente o mesmo.

# Comparando as principais categorias de incidentes por estado

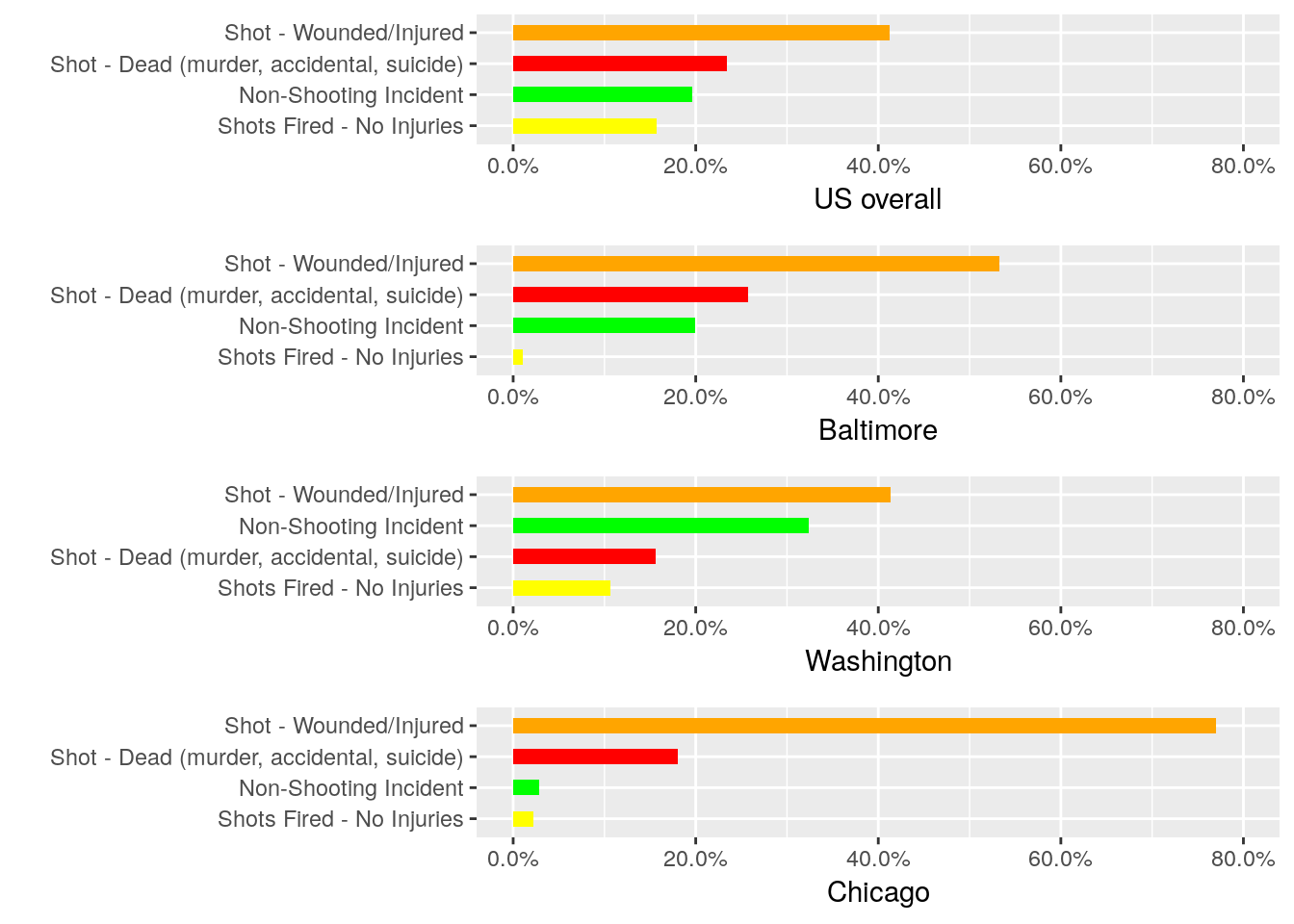
Primeiro de tudo, vou ver se há diferenças significativas ao comparar as categorias dos três estados com as taxas de incidentes mais altas com a distribuição nos EUA em geral.



Em relação aos estados com relativamente mais incidentes, podemos dizer que em comparação com as médias dos EUA:

* Em Louisiana, os incidentes são mais graves, com maiores porcentagens de incidentes com pessoas feridas e mortas.
* Em Delaware, houve mais incidentes com pessoas feridas, mas também menos incidentes com pessoas mortas
* No Alasca, os incidentes foram menos graves, com menos incidentes com feridos e mais incidentes sem tiro.

# Comparando as principais categorias de incidentes por cidade

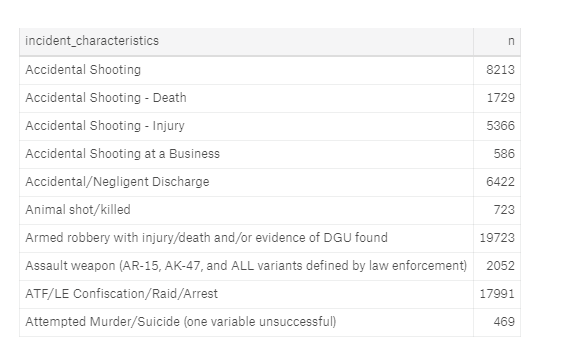


Em relação às cidades com relativamente mais incidentes, podemos dizer que:

* Em Baltimore, a porcentagem de incidentes com pessoas feridas é alta (20% acima da média dos EUA). Além disso, há muito poucos incidentes com tiros disparados e pessoas não feridas.
* Em Washington DC, apesar de ter uma média alta de incidentes por 100.000 incidentes, os incidentes são em média menos severos do que a média dos EUA. Relativamente, há mais incidentes sem filmagens e menos incidentes com vítimas mortais.
* Em Chicago, a porcentagem de incidentes com feridos é muito alta. Por outro lado, a porcentagem de incidentes com vítimas mortais é um pouco menor do que a média dos EUA.

# Oportunidades oferecidas pelas categorias menores de incidentes

As categorias de incidentes oferecem muitas informações e, na verdade, 110 categorias são registradas no banco de dados. Abaixo, estou exibindo as primeiras 10 categorias em ordem alfabética.



# Subcategorias dos incidentes mortais

A categoria geral "Shot - Dead (assassinato, acidental, suicídio)", por exemplo, consiste em várias subcategorias.

Presumo que assassinato / suicídio significa que a pessoa que cometeu suicídio, mas antes matou outras pessoas. Se eu tiver tempo, verificarei se esta lista de subcategorias está completa e inclua uma visualização. Outra subcategoria que vale a pena investigar é “Suicídio”, como por exemplo o artigo sobre o Alasca sugere uma alta taxa de suicídio neste estado.

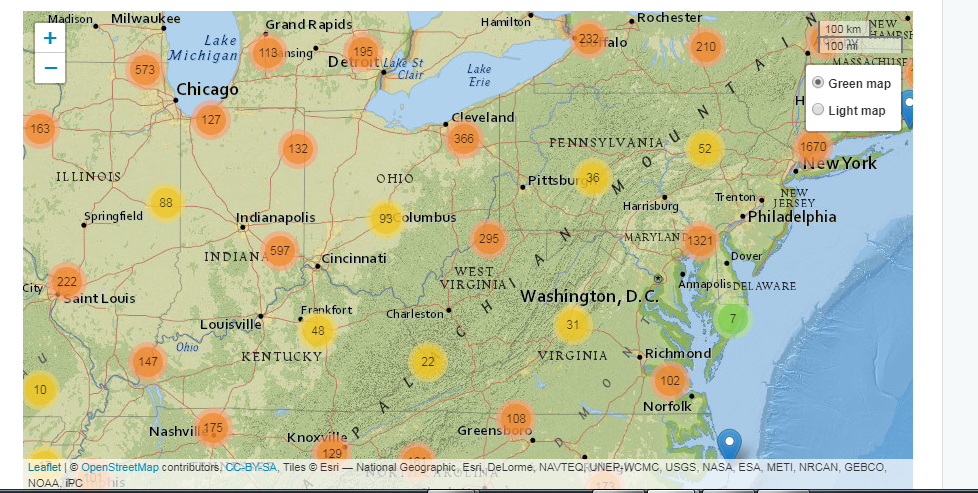
# Terrorismo, envolvimento de gangues e envolvimento de drogas

Outras subcategorias que podem ser bem visualizadas com os mapas de folhetos são o envolvimento com terrorismo, gangues e drogas.

### Envolvimento com drogas

Eu escolhi primeiro fazer um mapa do envolvimento com drogas, pois isso me obriga a usar o confronto de incidentes para tornar o mapa legível. Além disso, depois de ter experimentado com os níveis de zoom, decidi ampliar a área de Baltimore / Washington DC, já que, especialmente, Baltimore tinha muitos incidentes relacionados a drogas. Uma busca rápida na Internet me informou que isso é correto, veja também [Baltimore é a capital da heroína dos EUA](https://abcnews.go.com/US/story?id=92699&page=1).

Claro, você pode diminuir o zoom, mas isso tem a desvantagem de as áreas se tornarem tão grandes que várias cidades são agregadas. Passar pelo mapa enquanto mantém o nível de zoom igual dá a melhor visão geral do nível de "cidade" na minha opinião. Clicar nos clusters aproxima automaticamente essa área específica e divide um cluster em clusters menores. Você pode continuar fazendo isso até ver todos os incidentes individuais.

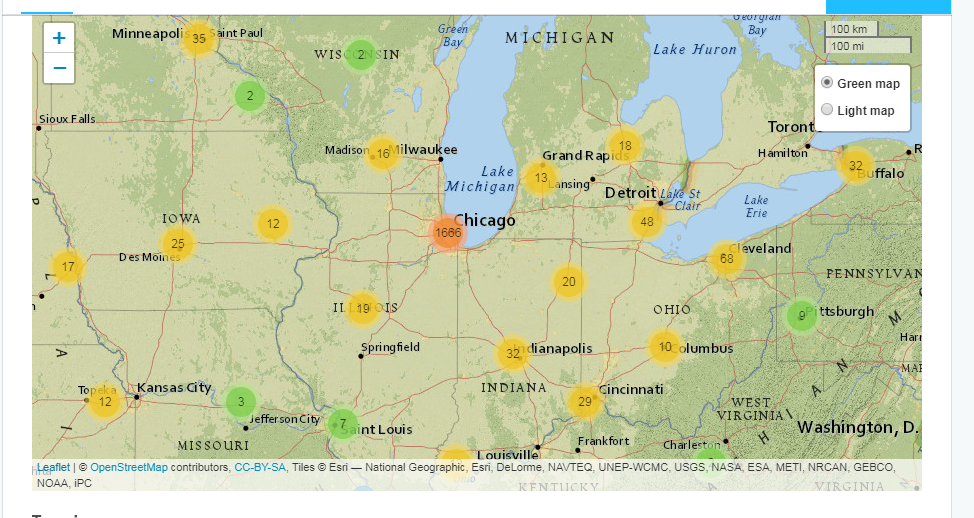


# Envolvimento de gangues

O mapa nesta seção é ampliado em Chicago, já que esta cidade tem, de longe, o maior número de incidentes com envolvimento de gangues. Por favor, também dê uma olhada neste artigo: [As 6 cidades mais infestadas de gangues na América](https://www.therichest.com/rich-list/most-shocking/the-most-gang-infested-cities-in-america/)

Nos últimos anos, Chicago deu um salto dramático para superar Los Angeles como a capital das gangues dos Estados Unidos, com um escalonamento de 150.000 membros de gangues e se tornou uma das cidades mais violentas do país.

Clicar nos clusters aproxima automaticamente essa área específica e divide um cluster em clusters menores. Você pode continuar fazendo isso até ver todos os incidentes individuais.



# Terrorismo

