Applied Problem Set 3

Guillermo Antonio Trefogli Wong & Earnest Salgado

28/05/2021

```
library(tidyverse)
library(lubridate)
library(tidycensus)
library(tmap)
library(ggrepel)
knitr::opts_chunk$set(fig.width=6, fig.height=3)
```

This submission is my work alone and complies with the 30535 integrity policy.

Add your initials to indicate your agreement: **GATW**

Add your collaborators: **ES**

Late coins used this pset: 0. Late coins left: 4.

Download the data for this Pset from here (link: https://www.kaggle.com/jameslko/gun-violence-data). Just click on Download and it should download automatically to your computer. Be careful when pushing back your work into github, this file is over 100MG so you won't be able to push your pset if the data is still there. This data set contains more than 260k gun violence incidents, with detail information about each incident. The data is from gunviolencearchive.org.

In the questions where you are asked to produce a graph or a map. Make sure that:

1. you add an informative title
2. your axis have informative names
3. graphs are not the default color (whichever you pick is fine)

Points will be taken off if your graphs and maps do not follow this.

1 Load the data and first glimpse (5 pts)

1. Load the data and show how many rows and columns it has

The data contains 239,677 rows and 29 columns:

```
gun_df <- read.csv("gun-violence-data_01-2013_03-2018.csv")
glimpse(gun_df)
```

```
## Rows: 239,677
## Columns: 29
## $ incident_id                <int> 461105, 460726, 478855, 478925, 478959, 47~
```

```
## $ date                         <chr> "2013-01-01", "2013-01-01", "2013-01-01", ~
## $ state                        <chr> "Pennsylvania", "California", "Ohio", "Col~
## $ city_or_county               <chr> "Mckeesport", "Hawthorne", "Lorain", "Auro~
## $ address                      <chr> "1506 Versailles Avenue and Coursin Street~
## $ n_killed                     <int> 0, 1, 1, 4, 2, 4, 5, 0, 0, 1, 1, 1, 2, 0, ~
## $ n_injured                    <int> 4, 3, 3, 0, 2, 0, 0, 5, 4, 6, 3, 3, 3, 5, ~
## $ incident_url                 <chr> "http://www.gunviolencearchive.org/inciden~
## $ source_url                   <chr> "http://www.post-gazette.com/local/south/2~
## $ incident_url_fields_missing <chr> "False", "False", "False", "False", "False~
## $ congressional_district       <int> 14, 43, 9, 6, 6, 1, 1, 2, 9, 7, 3, 1, 3, 1~
## $ gun_stolen                   <chr> "", "", "0::Unknown||1::Unknown", "", "0::~
## $ gun_type                     <chr> "", "", "0::Unknown||1::Unknown", "", "0::~
## $ incident_characteristics     <chr> "Shot - Wounded/Injured||Mass Shooting (4+~
## $ latitude                     <dbl> 40.3467, 33.9090, 41.4455, 39.6518, 36.114~
## $ location_description         <chr> "", "", "Cotton Club", "", "", "Fairmont T~
## $ longitude                    <dbl> -79.8559, -118.3330, -82.1377, -104.8020, ~
## $ n_guns_involved              <int> NA, NA, 2, NA, 2, NA, 2, NA, NA, NA, 1, 1,~
## $ notes                        <chr> "Julian Sims under investigation: Four Sho~
## $ participant_age              <chr> "0::20", "0::20", "0::25||1::31||2::33||3:~
## $ participant_age_group        <chr> "0::Adult 18+||1::Adult 18+||2::Adult 18+|~
## $ participant_gender           <chr> "0::Male||1::Male||3::Male||4::Female", "0~
## $ participant_name             <chr> "0::Julian Sims", "0::Bernard Gillis", "0:~
## $ participant_relationship     <chr> "", "", "", "", "3::Family", "", "5::Famil~
## $ participant_status           <chr> "0::Arrested||1::Injured||2::Injured||3::I~
## $ participant_type             <chr> "0::Victim||1::Victim||2::Victim||3::Victi~
## $ sources                      <chr> "http://pittsburgh.cbslocal.com/2013/01/01~
## $ state_house_district         <int> NA, 62, 56, 40, 62, 72, 10, 93, 11, NA, 28~
## $ state_senate_district        <int> NA, 35, 13, 28, 27, 11, 14, 5, 7, 44, 10, ~
```

```
head(gun_df)
```

```
##   incident_id       date          state city_or_county
## 1      461105 2013-01-01   Pennsylvania     Mckeesport
## 2      460726 2013-01-01     California      Hawthorne
## 3      478855 2013-01-01           Ohio         Lorain
## 4      478925 2013-01-05       Colorado         Aurora
## 5      478959 2013-01-07 North Carolina     Greensboro
## 6      478948 2013-01-07       Oklahoma          Tulsa
##                                    address n_killed n_injured
## 1 1506 Versailles Avenue and Coursin Street        0         4
## 2              13500 block of Cerise Avenue        1         3
## 3                     1776 East 28th Street        1         3
## 4          16000 block of East Ithaca Place        4         0
## 5                 307 Mourning Dove Terrace        2         2
## 6                6000 block of South Owasso        4         0
##                                     incident_url
## 1 http://www.gunviolencearchive.org/incident/461105
## 2 http://www.gunviolencearchive.org/incident/460726
## 3 http://www.gunviolencearchive.org/incident/478855
## 4 http://www.gunviolencearchive.org/incident/478925
## 5 http://www.gunviolencearchive.org/incident/478959
## 6 http://www.gunviolencearchive.org/incident/478948
##
## 1 http://www.post-gazette.com/local/south/2013/01/17/Man-arrested-in-New-Year-s-Eve-shooting-in-McKe
```

```
## 2                                                                    http://www.dailybulletin.com/article/
## 3                                    http://chronicle.northcoastnow.com/2013/02/14/2-men-indicted-in-ne
## 4                                  http://www.dailydemocrat.com/20130106/aurora-shootout-killer-was-frene
## 5                                           http://www.journalnow.com/news/local/article_d4c723e8-5a0f-1
## 6                       http://usnews.nbcnews.com/_news/2013/01/07/16397584-police-four-women-found-dead-in
##   incident_url_fields_missing congressional_district             gun_stolen
## 1                       False                      14
## 2                       False                      43
## 3                       False                       9 0::Unknown||1::Unknown
## 4                       False                       6
## 5                       False                       6 0::Unknown||1::Unknown
## 6                       False                       1
##                 gun_type
## 1
## 2
## 3 0::Unknown||1::Unknown
## 4
## 5 0::Handgun||1::Handgun
## 6
##
## 1                                        Shot - Wounded/Injured||Mass Shooting (4+ victims injured or
## 2                                                                                    Shot - Woun
## 3
## 4 Shot - Dead (murder, accidental, suicide)||Officer Involved Incident||Officer Involved Shooting - s
## 5
## 6                       Shot - Dead (murder, accidental, suicide)||Home Invasion||Home Invasion - Resid
##   latitude location_description longitude n_guns_involved
## 1  40.3467                      -79.8559              NA
## 2  33.9090                     -118.3330              NA
## 3  41.4455          Cotton Club  -82.1377               2
## 4  39.6518                     -104.8020              NA
## 5  36.1140                      -79.9569               2
## 6  36.2405     Fairmont Terrace  -95.9768              NA
##
## 1                                                                          Julian Sims under investig
## 2                                                                       Four Shot; One Killed; Unident
## 3
## 4
## 5 Two firearms recovered. (Attempted) murder suicide - both succeeded in fulfilling an M/S and did no
## 6
##                     participant_age
## 1                          0::20
## 2                          0::20
## 3 0::25||1::31||2::33||3::34||4::33
## 4       0::29||1::33||2::56||3::33
## 5       0::18||1::46||2::14||3::47
## 6       0::23||1::23||2::33||3::55
##                                                                participant_age_group
## 1              0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+||4::Adult 18+
## 2                          0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+
## 3              0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+||4::Adult 18+
## 4                          0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+
## 5                         0::Adult 18+||1::Adult 18+||2::Teen 12-17||3::Adult 18+
## 6 0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+||4::Adult 18+||5::Adult 18+
```

```
##                                         participant_gender
## 1                       0::Male||1::Male||3::Male||4::Female
## 2                                                0::Male
## 3                 0::Male||1::Male||2::Male||3::Male||4::Male
## 4                       0::Female||1::Male||2::Male||3::Male
## 5                     0::Female||1::Male||2::Male||3::Female
## 6 0::Female||1::Female||2::Female||3::Female||4::Male||5::Male
##                                                                                  participar
## 1                                                                                  0::Julia
## 2                                                                                0::Bernard
## 3                     0::Damien Bell||1::Desmen Noble||2::Herman Seagers||3::Ladd Tate Sr||4::Tallis
## 4                      0::Stacie Philbrook||1::Christopher Ratliffe||2::Anthony Ticali||3::Sonny Ar
## 5      0::Danielle Imani Jameison||1::Maurice Eugene Edmonds, Sr.||2::Maurice Edmonds II||3::Sandra
## 6 0::Rebeika Powell||1::Kayetie Melchor||2::Misty Nunley||3::Julie Jackson||4::James Poore||5::Cedri
##   participant_relationship
## 1
## 2
## 3
## 4
## 5               3::Family
## 6
##                                                          participant_status
## 1                            0::Arrested||1::Injured||2::Injured||3::Injured||4::Injured
## 2                                          0::Killed||1::Injured||2::Injured||3::Injured
## 3 0::Injured, Unharmed, Arrested||1::Unharmed, Arrested||2::Killed||3::Injured||4::Injured
## 4                                            0::Killed||1::Killed||2::Killed||3::Killed
## 5                                           0::Injured||1::Injured||2::Killed||3::Killed
## 6 0::Killed||1::Killed||2::Killed||3::Killed||4::Unharmed, Arrested||5::Unharmed, Arrested
##                                                                participant_type
## 1                    0::Victim||1::Victim||2::Victim||3::Victim||4::Subject-Suspect
## 2                    0::Victim||1::Victim||2::Victim||3::Victim||4::Subject-Suspect
## 3           0::Subject-Suspect||1::Subject-Suspect||2::Victim||3::Victim||4::Victim
## 4                              0::Victim||1::Victim||2::Victim||3::Subject-Suspect
## 5                              0::Victim||1::Victim||2::Victim||3::Subject-Suspect
## 6 0::Victim||1::Victim||2::Victim||3::Victim||4::Subject-Suspect||5::Subject-Suspect
##
## 1
## 2
## 3
## 4 http://denver.cbslocal.com/2013/01/06/officer-told-neighbor-standoff-gunman-was-on-meth-binge/||ht
## 5
## 6       http://www.kjrh.com/news/local-news/4-found-shot-inside-apartment-in-tulsa||http://www.cbsn
##   state_house_district state_senate_district
## 1                   NA                   NA
## 2                   62                   35
## 3                   56                   13
## 4                   40                   28
## 5                   62                   27
## 6                   72                   11
```

2. Explore the variables on your own (we do not want to see any code here) and write a short paragraph of what you find. Pay attention on which characteristics we have for each event.

Based on the description of the variables we can observe that the data contains a unique identification for

events associated with gun violence. The variables describe the events in terms of different characteristics, such as basic information of the event (date, location, number of people involved), the level of violence (number of killed or injured, number of guns, type of guns) the features of the people involved (age, gender, relationship) and other related information (electoral district, website url, notes, etc.).

3. Which variables you might have to format later? (just name them you do not have to do anything else right now)

Changes in terms of format and to tidy the data:

- We will probably need to change the type of some variables. For instance, "Date" is a character type and we will probably need it to be as date, or "participant_age" is a character and we will probably need it to be a factor.
- It could the case that some variables have redundant information than others, such as "participant_age" and "participant_age_group"
- It seems that, in order to analyze some variables, we will need to tidy in through pivot_longer or pivot_wider, because some are containing information for more than one observation. It maybe the case of tidying the data relative to "participant_gender" or "participant name" variables.
- Some missing values are explicit, other are implicit
- We will need to standardize the format of some values. For instance it is the case of "gun_type"
- Some variables have many information inside, so that if we want to analyze them we will probably need to modify them through separating they strings, earsing unhelpful information, etc.. For instance, it is the case of "incident characteristics" or "participant name"

2 Time related trends of gun violence (10 pts)

1. Make sure the data variable is in a Date format. If not transform it.

```
gun_date <- gun_df %>% mutate(date = ymd(date))
```

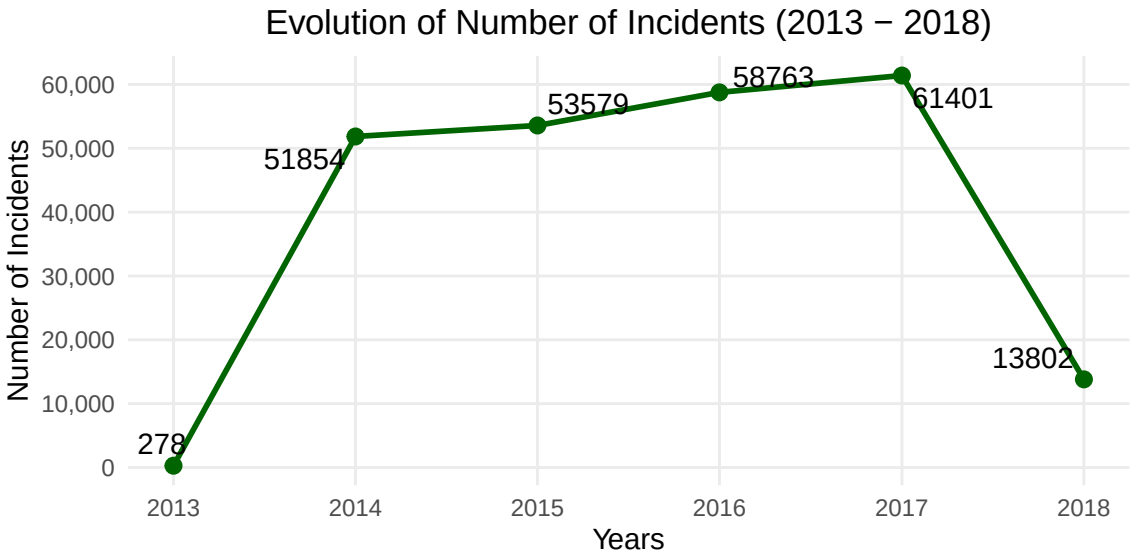
2. Is the number of incidents increasing by year? Show your result in a graph and write a short paragraph about what you are seeing. Add the count for each year as an annotation label.

```
gun_year <- gun_date %>%
  mutate(year = floor_date(date, unit = "year"))

df_year <- gun_year %>%
  group_by(year, incident_id) %>%
  summarise(n = n()) %>%
    group_by(year) %>%
    summarise(n_inci=n())
```

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```

```
df_year %>%
  ggplot(aes(year, n_inci)) +
  geom_line(size = 1, color = "darkgreen") + geom_point(size = 2.5, color = "darkgreen") +
  labs(title = "Evolution of Number of Incidents (2013 - 2018)",
       x = "Years",
       y = "Number of Incidents") +
```

```
  geom_text_repel(aes(label = n_inci)) +
  scale_y_continuous(labels = scales::comma, breaks = seq(0, 65000, by = 10000)) +
  scale_x_date(date_breaks = "1 year", date_labels ="%Y" ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5), panel.grid.minor = element_blank())
```



The number of incidents increased dramatically from 2013 to 2014. Then, the increase continue, at a low rate, up to 2017. From 2017 to 2018, this number fall substantially. However, it reached a number still roughly five times that of the year 2013.

Important to note, it caught our attention the tails of the plot, 2013 and 2018.

```
summary(gun_date$date)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2013-01-01" "2015-03-07" "2016-04-05" "2016-03-13" "2017-04-03" "2018-03-31"
```

After observing the data, we noticed that in the case of 2018 the low numbers are explained, in part, because of the difference in the period covered for that year. Then, for further analysis, one caveat at this point will be not to consider 2018 when doing comparisons at years level.

On the other hand, regarding 2013, the values are in fact very low, in comparison to following years. Then, one caveat here will be verify the quality of the data reported, to be sure that these numbers are different due to actual differences in incidents happening in the city, or because of differences in capacity of the organization to report the data.

3. Is there a particular violent month? Answer this question with a graph. Be careful: should you use the whole data to answer this question? Add the count for each month as an annotation label.

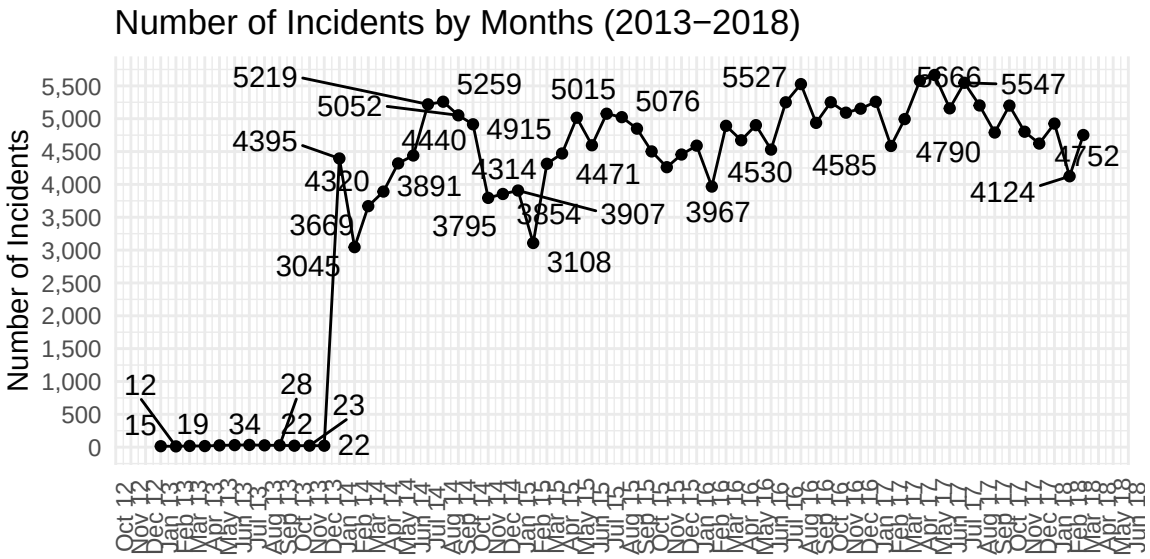
To perform this analysis, it would be more informative to exclude the year 2013 because there is a huge difference between the number of incidents registered in that year from those of the following years. To better understand this situation, it would be helpful to visualize the pattern in a plot, as follows:

```
gun_month <- gun_date %>% mutate(month = floor_date(date, unit = "month"))

# glimpse(df_month)

df_month <- gun_month %>%
  group_by(month, incident_id) %>%
  summarise(n = n()) %>%
    group_by(month) %>%
    summarise(n_inci=n())
```

```
## `summarise()` has grouped output by 'month'. You can override using the `.groups` argument.
```

```
df_month %>%
  ggplot(aes(month, n_inci)) +
  geom_line() + geom_point() +
  labs(title = "Number of Incidents by Months (2013-2018)",
       x = "Month",
       y = "Number of Incidents") +
  geom_text_repel(aes(label = n_inci)) +
  scale_x_date(NULL, date_labels = "%b %y", breaks = "month") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_y_continuous(labels = scales::comma, breaks = seq(0, 5666, by = 500))
```

```
## Warning: ggrepel: 28 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Then, plotting the average number of incidents by month excluding year 2013:

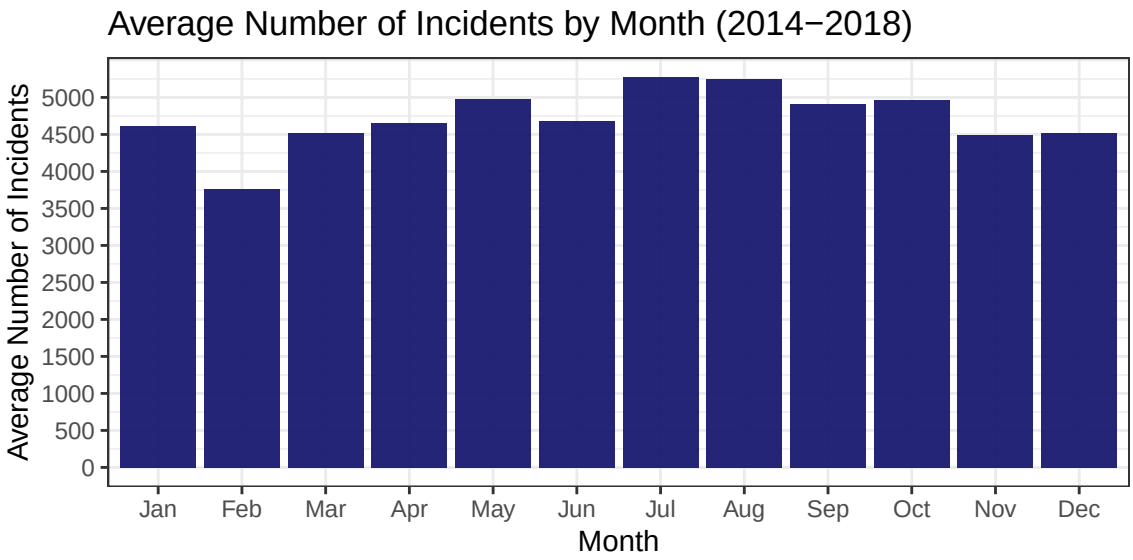
```
gun_bymonth <- gun_year %>%
  filter(year(date) != 2013) %>%
  mutate(datemonth = floor_date(date, unit = "month"),
```

```
                                    month = month(date, label = TRUE))

(gun_bymonth_chart <- gun_bymonth %>%
  group_by(datemonth, month) %>%
  summarise(n = n())%>%
  group_by(month) %>%
  summarise(avg_month = mean(n)))
```

```
## `summarise()` has grouped output by 'datemonth'. You can override using the `.groups` argument.
```

```
## # A tibble: 12 x 2
##    month avg_month
##    <ord>     <dbl>
##  1 Jan       4615.
##  2 Feb       3766.
##  3 Mar       4524.
##  4 Apr       4653.
##  5 May       4976.
##  6 Jun       4681
##  7 Jul       5273
##  8 Aug       5253
##  9 Sep       4907
## 10 Oct       4967
## 11 Nov       4488.
## 12 Dec       4521
```

```
gun_bymonth_chart %>%
  ggplot(aes(month, avg_month)) +
  geom_bar(stat = "identity", alpha = 0.95, fill = "midnightblue") +
  labs(title = "Average Number of Incidents by Month (2014-2018)",
       x = "Month",
       y = "Average Number of Incidents") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(breaks = seq(0, 5273, by = 500))
```



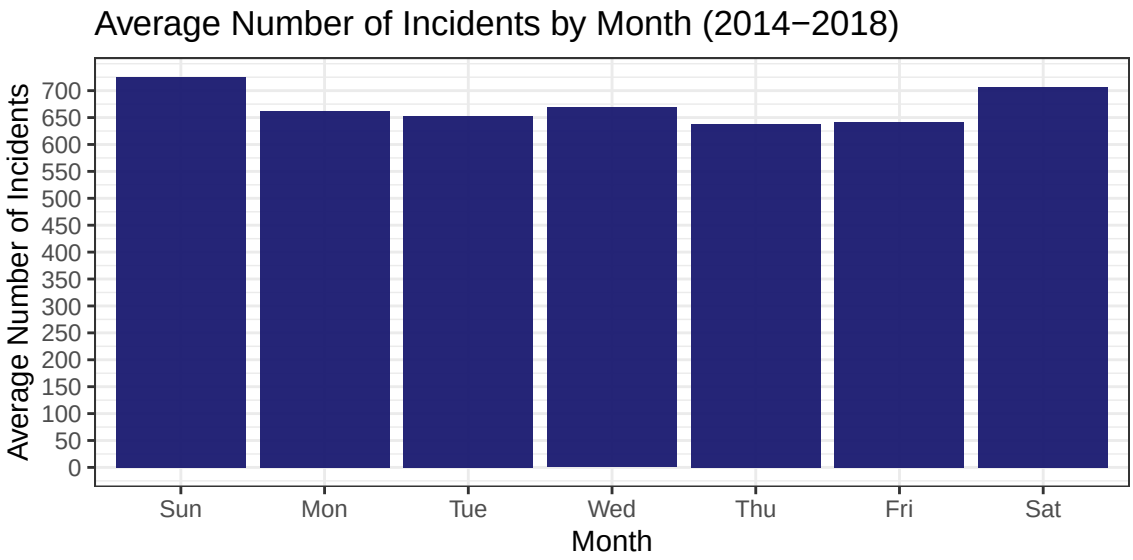
It can be notice that, on average, July and August are the months with the highest number of incidents.

4. Is there a particular violent day of the week? Add the count for each day as an annotation label.

```
gun_byweek <- gun_bymonth %>% mutate(datemonth = floor_date(date, unit = "month"),
                                     weekday = wday(date, label = TRUE))

(gun_byweek_chart <- gun_byweek %>%
  group_by(datemonth, weekday) %>%
  summarise(n = n())%>%
  group_by(weekday) %>%
  summarise(avg_weekday = mean(n)))
```

```
## `summarise()` has grouped output by 'datemonth'. You can override using the `.groups` argument.

## # A tibble: 7 x 2
##   weekday avg_weekday
##   <ord>         <dbl>
## 1 Sun            725.
## 2 Mon            661.
## 3 Tue            652.
## 4 Wed            669.
## 5 Thu            638.
## 6 Fri            642.
## 7 Sat            707.
```

```
gun_byweek_chart %>%
  ggplot(aes(weekday, avg_weekday)) +
  geom_bar(stat = "identity", alpha = 0.95, fill = "midnightblue") +
  labs(title = "Average Number of Incidents by Month (2014-2018)",
       x = "Month",
       y = "Average Number of Incidents") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw() +
  scale_y_continuous(breaks = seq(0, 724.9, by = 50))
```



In terms of weekday, it seems that it is Sunday the with the highest number of incidents.

5. Write a short paragraph with your findings on the number of incidents by year, month, and weekday.

Based on the previous plots and information, our takeaways are the following ones:

- In terms of years, the yearly average incidents is 56399.25, only considering the years 2014-2017. The yearly number of incidents have been increasing from 2014 to 2007 at a average yearly number of 3182.
- In terms of months, July and August, the ones related with summer season, are the ones with the highest number of incidents. The difference however is not very high: the average of incidents of these two months is 5263, while the average of the rest of months is 4609.745.
- In terms of weekday, Sunday is the one with the highest number of incidents, followed by Saturday. Then, weekends are the ones with more incidents. On average, a day in the weekends has 724.902 incidents, while the average of the rest of weekdays is 652.5294.

3 Characteristics of Gun violence incidents (20 points)

WARNING!! some variables in this data set need to be manipulated before you try to plot them. It might be a good idea to go back and check your notes on Strings, Tidy, and Joins.

1. What is the average number of guns involved in an incident?

For this question we perform mean() on the n_guns_involved variable and calculate 1.372442 guns on average.

```
mean(gun_df$n_guns_involved, na.rm = TRUE)
```

```
## [1] 1.372442
```

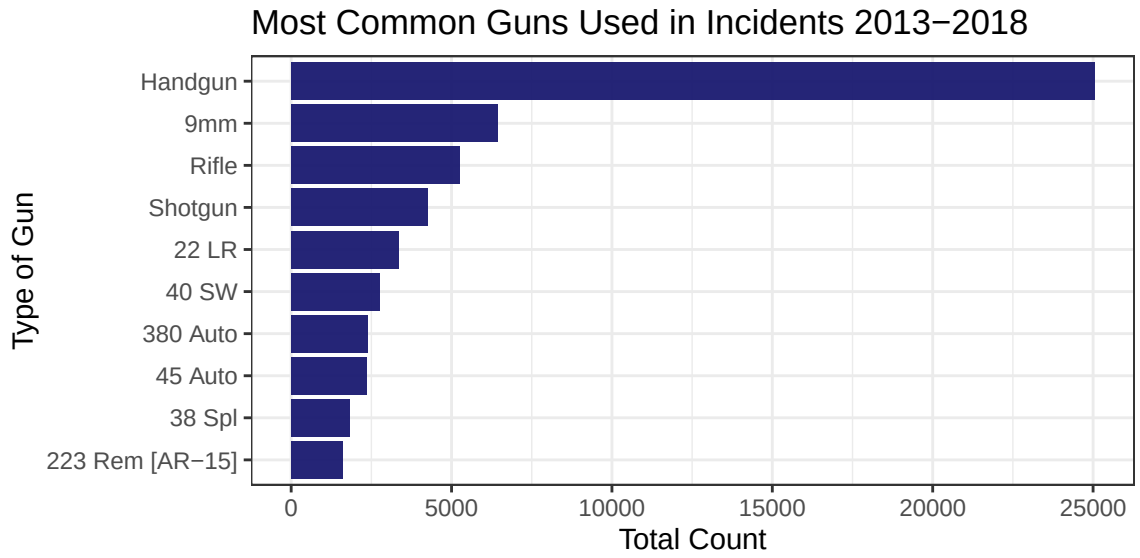
```
# 1.372442 guns involved in each incident
```

2. Which type of guns are more commonly used? Use a plot to show your answer. Do not show Unknown or missing values on your plot.

Manipulation is needed to create a plot that is clear and tidy. Functions that are beneficial are separate_rows(), str_detect(), and str_replace. Although there of course could be other functions to manipulate the data.

Looking at the data in the plot, we found Handguns to be the most common gun type, which was almost five times more common as the second-most common gun, the 9mm.

```
type <- gun_df %>%
  select(incident_id, date, gun_type)

type2 <- type %>%
  separate_rows(gun_type, sep = "\\|")

type2[type2 == ""] = NA

type3 <- type2 %>%
  filter(!str_detect(gun_type, 'Unknown')) %>%
  mutate(gun_type = str_replace(gun_type, '[0-9]+::', "")) %>%
  group_by(gun_type) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  head(10)

type3 %>%
  ggplot(aes(x = fct_reorder(gun_type, n), y = n)) +
  geom_bar(stat = "identity", alpha = 0.95, fill = "midnightblue") +
  labs(title = "Most Common Guns Used in Incidents 2013-2018",
       x = "Type of Gun",
       y = "Total Count") +
  theme_bw() +
  coord_flip()
```



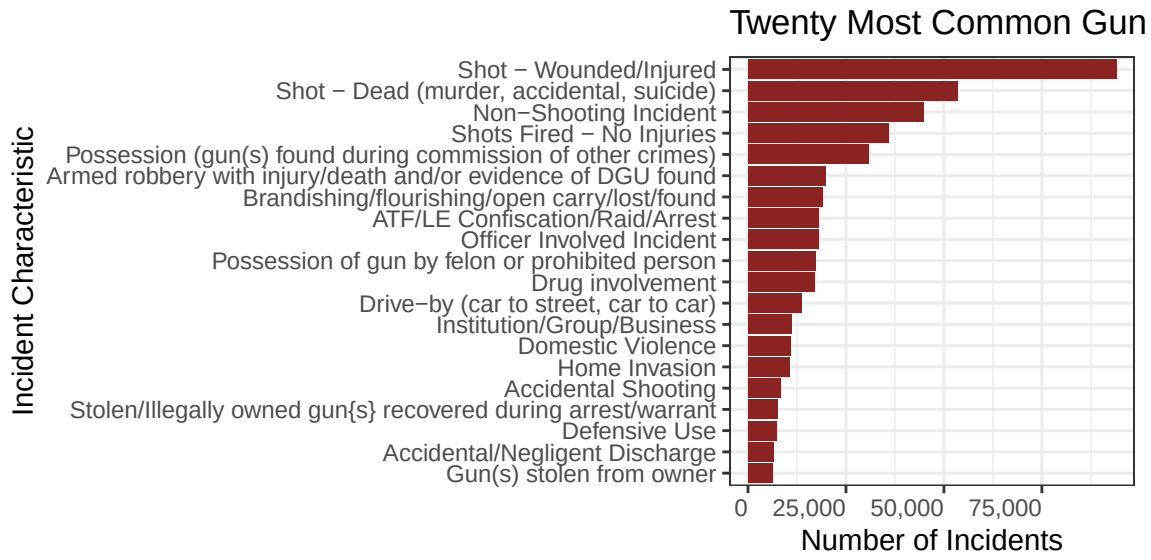
3. Explore the incident_characteristics variable. What is this variable telling us? Why are there multiple characteristics for each incident?

The incident_characteristics variable tells us what happened during this gun-related incident, whether participants were shot or not, wounded/injured, or killed. It also describes the incident type, which can be suicide, drive-by's, or mass shootings, for example. It also quickly describes the geographical location or setting of the incident (e.g. armed-robbery, police-involved shooting, domestic violence). There seems to be multiple characteristics for each incident to better record different aspects of the incident, and to provide information when there are multiple participants and/or reasons for the incident.

4. Show in a plot which are the most common incident_characteristics cap your graph at 20 (i.e. just show the 20 most common incident characteristics)

```
incidents <- gun_df %>%
  select(incident_id, date, incident_characteristics)

incidents2 <- incidents %>%
  separate_rows(incident_characteristics, sep = "\\|")

incidents2[incidents2 == ""] = NA

top20incidents <- incidents2 %>%
  filter(incident_characteristics != is.na(.)) %>%
  group_by(incident_characteristics) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  head(20)
```

```
top20incidents %>%
  ggplot(aes(x = fct_reorder(incident_characteristics, n), y = n)) +
  geom_bar(stat = "identity", fill = "brown4") +
  scale_y_continuous("Number of Incidents", labels = scales::comma) +
  scale_x_discrete("Incident Characteristic") +
  ggtitle("Twenty Most Common Gun Incidents") +
  coord_flip() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1))
```



4 Suspects Characteristics (30 points)

1. Explore the following variables and write a short paragraph of what they mean and how they are connected to each other: participant_age, participant_gender, participant_type, and participant_status. Do you see any technical difficulties for how these variables are coded? If so, explain them.

We explore and define the following variables: participant_age = age of each participant in an incident, participant_gender = either male or female for participants involved, participant_type = victims or subject/suspects in a particular incident, participant_status = characterizes how each participant was following the incident. There were many unique values for participant_status, "Killed", "Injured", "UnharmedArrested", "Arrested", "Unharmed", "InjuredArrested". "KilledArrested".

In terms of "technical difficulties", we observe these cells store multiple, discrete instances of data. The [:digit:] at the start of each value is presumably used to track the values accross from one "participant_" variable to the next. For example, one observation of participant_age contains only one age value, with four participant_gender values. One avenue for addressing these inconsistencies could involve performing counts for each delimited value. We manipulate these variables in order to correctly count.

2. What is the average of suspects and victims per incident?

```
suspects_victims <- gun_df %>%
  select(incident_id, participant_type) %>%
  separate_rows(participant_type, sep = "\\|")

suspects_victims[suspects_victims == ""] = NA

suspects_victims2 <- suspects_victims %>%
  filter(participant_type != is.na(.)) %>%
  mutate(participant_type = str_replace(participant_type, '[0-9]+::', ""),
         participant_type = str_replace(participant_type, '[0-9]+:', ""))

unique(suspects_victims2$participant_type)
```

```
## [1] "Victim"          "Subject-Suspect"
```

```
suspects_victims3 <- suspects_victims2 %>%
  group_by(incident_id, participant_type) %>%
  summarise(n = n())
```

```
## `summarise()` has grouped output by 'incident_id'. You can override using the `.groups` argument.
```

```
sus_vict_wider <- suspects_victims3 %>%
  pivot_wider(names_from = participant_type, values_from = n)

sus_vict_wider[is.na(sus_vict_wider)] <- 0

# pivot_wider
```

```
sum(sus_vict_wider$Victim) / length(sus_vict_wider$incident_id)
```

```
## [1] 0.898731
```

```r
sum(sus_vict_wider$`Subject-Suspect`) / length(sus_vict_wider$incident_id)
```

```
## [1] 0.9276071
```

```r
summary(sus_vict_wider$`Subject-Suspect`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.9276  1.0000 63.0000
```

```r
summary(sus_vict_wider$Victim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0000   1.0000   0.8987   1.0000 102.0000
```

3. Create a new data frame with just the suspects include the following variables: incident_id, participant_age, participant_gender. Just print the head() of this new data frame. (Hints: 1. Google what the function unnest does, is part of tidyr library, 2. review the tidy and joins materials)

```r
names(gun_df)
```

```
##  [1] "incident_id"                "date"
##  [3] "state"                      "city_or_county"
##  [5] "address"                    "n_killed"
##  [7] "n_injured"                  "incident_url"
##  [9] "source_url"                 "incident_url_fields_missing"
## [11] "congressional_district"     "gun_stolen"
## [13] "gun_type"                   "incident_characteristics"
## [15] "latitude"                   "location_description"
## [17] "longitude"                  "n_guns_involved"
## [19] "notes"                      "participant_age"
## [21] "participant_age_group"      "participant_gender"
## [23] "participant_name"           "participant_relationship"
## [25] "participant_status"         "participant_type"
## [27] "sources"                    "state_house_district"
## [29] "state_senate_district"
```

```r
suspects <- gun_df %>%
  select(incident_id, participant_age, participant_gender)

suspects_split <- suspects %>%
transform(participant_age = strsplit(participant_age, "\\|"),
          participant_gender = strsplit(participant_gender, "\\|")) %>%
  unnest(participant_age) %>%
  unnest(participant_gender)

suspects_split[suspects_split  == ""] = NA

suspects_split2 <- suspects_split %>%
  mutate(participant_age = str_replace(participant_age, '[0-9]+::', ""),
         participant_gender = str_replace(participant_gender, '[0-9]+::', "")) %>%
    filter(across(c(participant_age, participant_gender), ~ !is.na(.x))) %>%
  head()
```

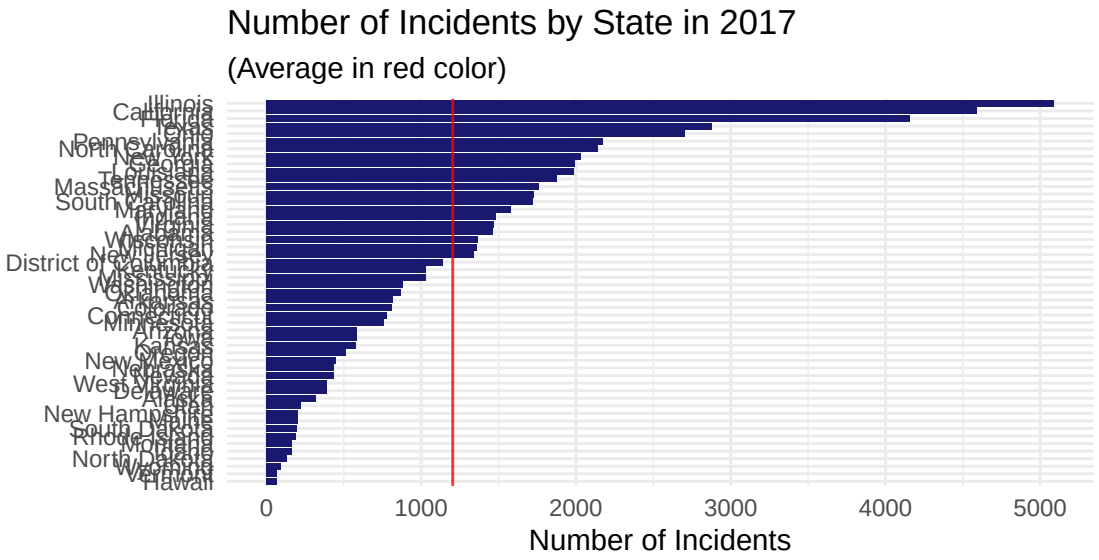
4. Show the distribution of suspects age, crop your plot if you find any suspect over the age of 100
5. What percentage of suspects are male (exclude missing values)?
6. How many different status are there?
7. What percentage of all suspects got arrested? Be careful for some suspects there are more than 1 categories.

5 Geographic variation (15 points)

1. What was the state with more incidents in 2017? Use a graph to answer this question.

```
incidents17 <- gun_month %>%
  filter(year(date) == 2017) %>%
  group_by(state) %>%
  summarise(n=n()) %>%
  arrange(desc(n))

avg <- mean(incidents17$n)

incidents17 %>%
  ggplot(aes(x = fct_reorder(state, n), y = n)) +
  geom_col(stats = "identity", position = "dodge", fill = "midnightblue") +
  geom_hline(aes(yintercept = avg), color="red", alpha = 0.8) +
  coord_flip() +
  labs(title = "Number of Incidents by State in 2017",
       subtitle = "(Average in red color)",
       x = "",
       y = "Number of Incidents") +
  theme(plot.title = element_text(hjust = 0.5)) +
    theme_minimal()
```

```
## Warning: Ignoring unknown parameters: stats
```



2. Use your census API to get population by state, remember to also download the geometry we will use it later. Re-do your previous plot but adjusting by population, i.e. incidents by a 100,000 inhabitants.

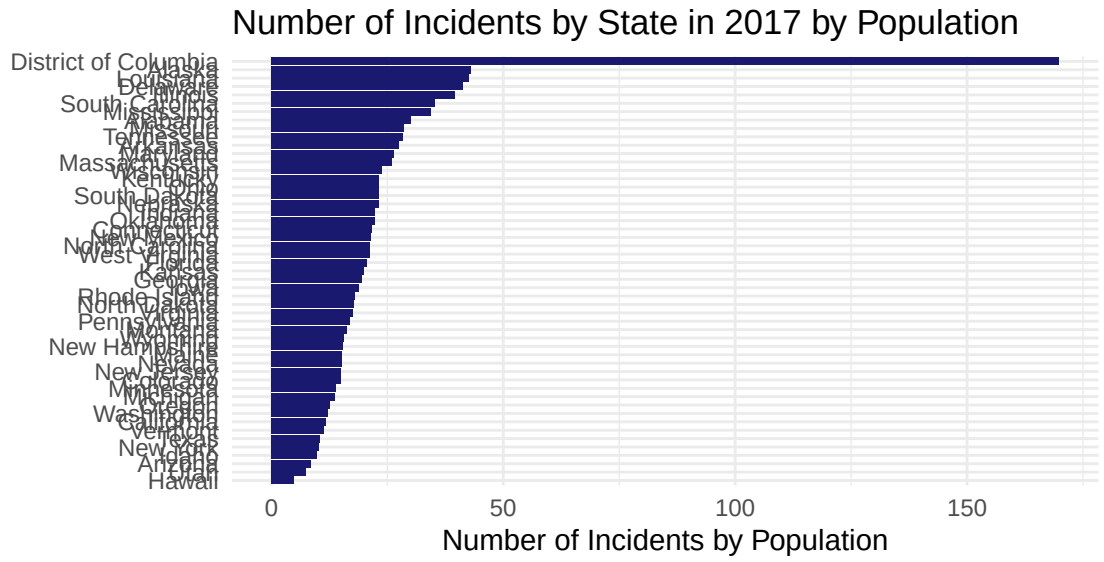
```
census_data.sp <- get_acs(
  geography = "state",
  variables = c(population = "B01003_001"),
  year = 2017,
  geometry = TRUE)
```

```
## Getting data from the 2013-2017 5-year ACS
```

```
## Downloading feature geometry from the Census website.  To cache shapefiles for use in future sessions
```

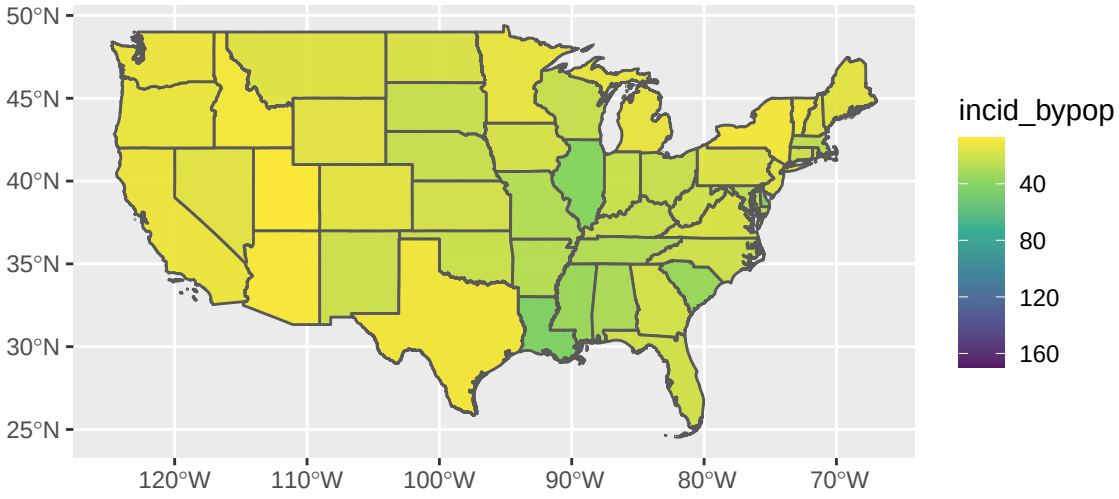
```
##   |                                                                              |
```

```
incidents17_bypop <- left_join(census_data.sp, incidents17, by = c("NAME" = "state")) %>%
  mutate(incid_bypop = n*100000/estimate)

incidents17_bypop %>%
  filter(NAME != "Puerto Rico") %>%
  ggplot(aes(x = fct_reorder(NAME, incid_bypop ), y = incid_bypop )) +
  geom_col( position = "dodge", fill = "midnightblue") +
  coord_flip() +
  labs(title = "Number of Incidents by State in 2017 by Population",
       x = "",
       y = "Number of Incidents by Population") +
    theme_minimal()
```



3. Show the results from your previous plot in a map

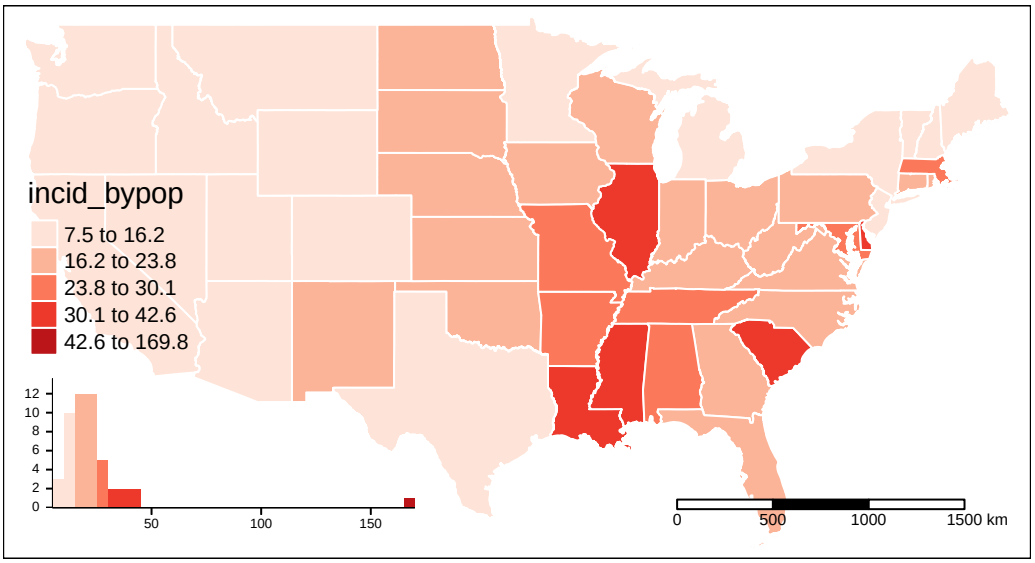
```
incidents17_bypop_us <-  incidents17_bypop%>%
  filter(NAME != "Alaska") %>%
  filter(NAME != "Hawaii") %>%
  filter(NAME != "Puerto Rico")

incidents17_bypop_us %>%
  ggplot(aes(fill = incid_bypop)) +
  geom_sf() +
  scale_fill_viridis_c(trans = "reverse", alpha = 0.9)
```




```
tm_shape(incidents17_bypop_us) +
  tm_fill(col = "incid_bypop", style = "jenks",  palette ="Reds", legend.hist=TRUE) +
  tm_borders(col = "white") +
  tm_scale_bar()
```

```
## Warning: package 'sf' was built under R version 4.0.5
```

```
## Linking to GEOS 3.9.0, GDAL 3.2.1, PROJ 7.2.1
```




```
## [1] "#FEE4D8" "#FCB499" "#FB795A" "#ED392B" "#BB1419"
```

6 Mass shootings (20 points)

1. Create a new data frame of mass shootings.

```
mass_shooting <-  gun_month %>%
  filter(str_detect(incident_characteristics, "Mass Shooting") == TRUE)
```

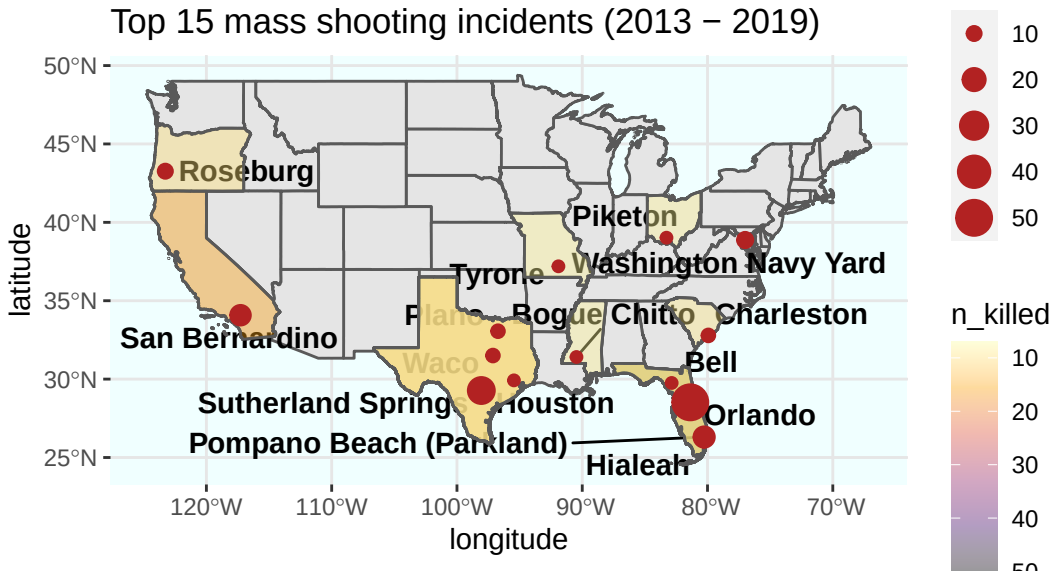
2. Show the top 15 incidents by number of victims in a map as points. The points should be proportional to the number of victims. Follow the map slides to make the map. Here are the coordinates that will give you the continental US out of the world map. Extra credit if you add state lines and change the theme of the map.

```
mass_shooting_t15 <- mass_shooting %>%
  arrange(desc(n_killed)) %>%
  head(15)

ms_t15.sp <- right_join(census_data.sp, mass_shooting_t15,  by = c("NAME" = "state"))%>%
  filter(NAME != "Alaska") %>%
  filter(NAME != "Hawaii") %>%
  filter(NAME != "Puerto Rico") %>%
  rename(state = NAME)

glimpse(ms_t15.sp)
```

```
## Rows: 15
## Columns: 35
## $ GEOID                    <chr> "45", "06", "41", "39", "48", "48", "48", ~
## $ state                    <chr> "South Carolina", "California", "Oregon", ~
## $ variable                 <chr> "population", "population", "population", ~
```

```
## $ estimate                    <dbl> 4893444, 38982847, 4025127, 11609756, 2741~
## $ moe                         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ incident_id                 <int> 359830, 456893, 423223, 545525, 980577, 34~
## $ date                        <date> 2015-06-17, 2015-12-02, 2015-10-01, 2016-~
## $ city_or_county              <chr> "Charleston", "San Bernardino", "Roseburg"~
## $ address                     <chr> "110 Calhoun Street", "1365 South Waterman~
## $ n_killed                    <int> 9, 16, 10, 8, 27, 9, 9, 8, 11, 8, 50, 17, ~
## $ n_injured                   <int> 0, 19, 9, 0, 20, 18, 1, 0, 3, 1, 53, 17, 0~
## $ incident_url                <chr> "http://www.gunviolencearchive.org/inciden~
## $ source_url                  <chr> "http://7online.com/news/charleston-shooti~
## $ incident_url_fields_missing <chr> "False", "False", "False", "False", "False~
## $ congressional_district      <int> 1, 31, 4, 2, 28, 17, 3, 18, 1, 3, 5, 22, 3~
## $ gun_stolen                  <chr> "0::Unknown", "0::Not-stolen||1::Not-stole~
## $ gun_type                    <chr> "0::45 Auto", "0::223 Rem [AR-15]||1::223 ~
## $ incident_characteristics    <chr> "Shot - Dead (murder, accidental, suicide)~
## $ latitude                    <dbl> 32.7876, 34.0758, 43.2628, 39.0201, 29.273~
## $ location_description        <chr> "Mother Emanuel AME Church", "Inland Regio~
## $ longitude                   <dbl> -79.9332, -117.2770, -123.2800, -83.2750, ~
## $ n_guns_involved             <int> 1, 4, 6, 1, 4, NA, 1, 1, 2, 1, 2, 2, NA, 1~
## $ notes                       <chr> "at least 8 dead, more inj Black church, w~
## $ participant_age             <chr> "0::41||1::26||2::74||3::54||4::46||5::70|~
## $ participant_age_group       <chr> "0::Adult 18+||1::Adult 18+||2::Adult 18+|~
## $ participant_gender          <chr> "0::Male||1::Male||2::Male||3::Female||4::~
## $ participant_name            <chr> "0::Rev. Clementa Pinckney||1::Tywanza San~
## $ participant_relationship    <chr> "", "", "", "8::Family", "", "", "", "8::F~
## $ participant_status          <chr> "0::Killed||1::Killed||2::Killed||3::Kille~
## $ participant_type            <chr> "0::Victim||1::Victim||2::Victim||3::Victi~
## $ sources                     <chr> "http://abc11.com/news/charleston-sc-polic~
## $ state_house_district        <int> 111, 40, 7, 90, 44, NA, 67, 140, NA, 53, 4~
## $ state_senate_district       <int> 42, 23, 4, 14, 21, 22, 8, 15, NA, 39, 12, ~
## $ month                       <date> 2015-06-01, 2015-12-01, 2015-10-01, 2016-~
## $ geometry                    <MULTIPOLYGON [°]> MULTIPOLYGON (((-79.50795 3..~
```

```
locations <- ms_t15.sp %>%
  select(state, city_or_county, latitude, longitude, n_killed)

locations.sp <- st_as_sf(locations, coords = c("longtitude", "latitude"),  remove = FALSE, crs = 4269,
                         agr = "constant")

ggplot(incidents17_bypop_us) +
  geom_sf() +
  geom_sf(data = locations.sp) +
    geom_text_repel(data = locations.sp, aes(x = longitude, y = latitude, label = city_or_county),
        fontface = "bold") +
  geom_sf(data = ms_t15.sp, aes(fill = n_killed))+
  scale_fill_viridis_c(trans = "reverse", alpha = 0.4, option = "B") +
  geom_point(data = ms_t15.sp, aes(x = longitude, y = latitude, size = n_killed), col = "Firebrick") +
  ggtitle("Top 15 mass shooting incidents (2013 - 2019)") + #VERIFY DATES
  theme(panel.grid.major = element_line(color = gray(0.9),
        size = 0.5), panel.background = element_rect(fill = "azure"))
```



3. Use the data set you create about suspects in previous sections to explore if there are differences between the suspects of mass shootings vs other types of incidents. Compare suspects age, gender, percentage of arrested, percentage of killed and number of shooters/suspects per incident. Write a short paragraph of your findings.

```
ms_suspects <-  gun_month %>%
  mutate(suspect = ifelse(str_detect(incident_characteristics, "Mass Shooting") == TRUE, "Suspect Mass S
  # I think we should put a pipe here to select the variables fot the analysis, something like this:
  # %>% select(incident_id, participant_age, participant_gender, participant_status, participant_type, 
```