Applied 3

SANDRA DAI & Qianyu Shao

12/01/2022

```
library(tidyverse)
library(lubridate)
library("maps")
library(sf)
library(rnaturalearth)
library(rnaturalearthdata)
```

1 Front matter

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

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

Add your collaborators: **QS**

Late coins used this pset: 1. Late coins left: 2.

2 1

2.1 Load the data and first glimpse

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

```
gun <- read_csv("/Users/sandradai/Downloads/gun-violence-data_01-2013_03-2018.csv")
```

```
## Rows: 239677 Columns: 29
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (18): state, city_or_county, address, incident_url, source_url, gun_sto...
## dbl   (9): incident_id, n_killed, n_injured, congressional_district, latitud...
## lgl   (1): incident_url_fields_missing
## date  (1): date
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
nrow(gun)
```

```
## [1] 239677
```

```
ncol(gun)
```

```
## [1] 29
```

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.

- the dataset has information about the description, location, participants in gun violence incidents in the U.S. from 2013 to 2018. The incident characteristics described the means and outcome of the gun violence crimes.

3. Which variables you might have to format later?

- incident_characteristics
- date
- participant_age, participant_age_group, participant_gender, participant_name, participant_status

3 2

3.1 Time ralated trends of gun violence

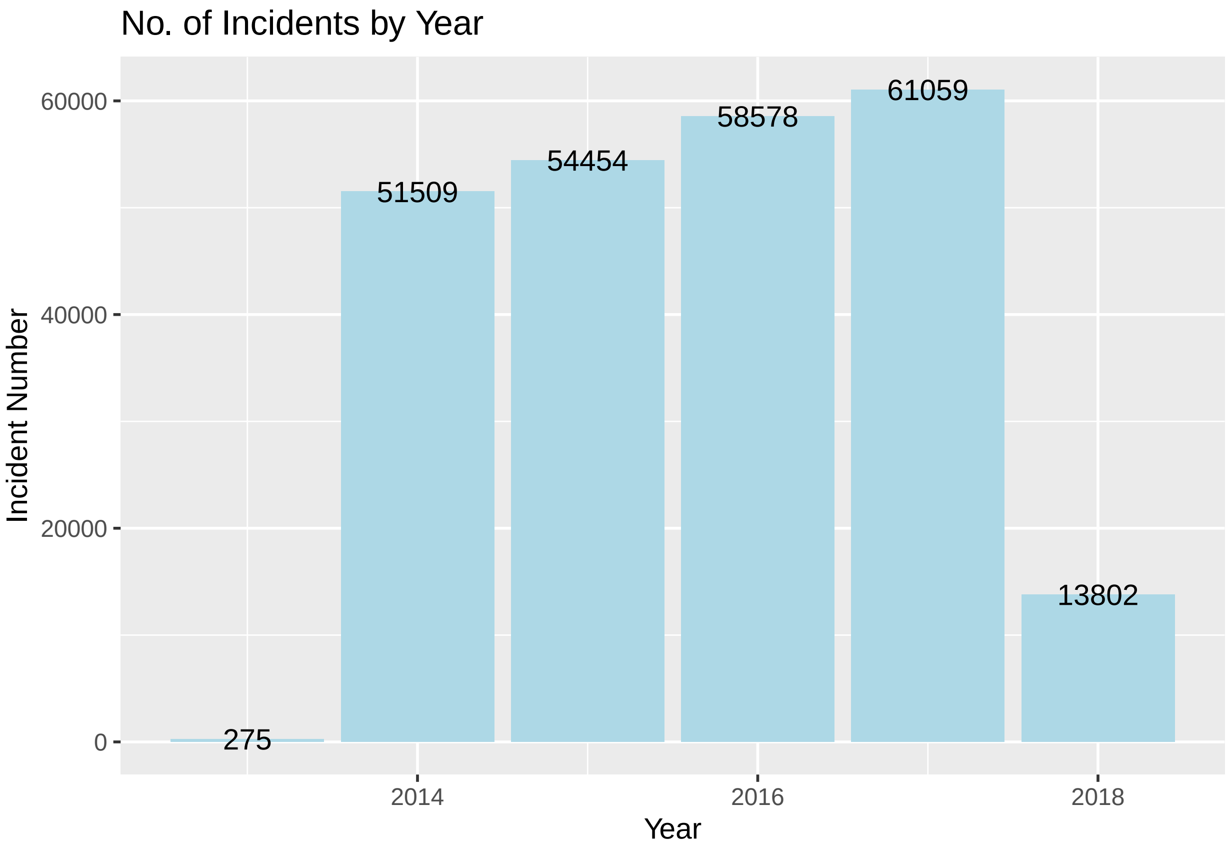
1. Make sure the date variable is in a Date format. If not transform it.

```
gun$date <- as.Date(gun$date, "%m/%d/%Y")
```

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.

- According to the graph, the number of incidents increased during 2013 to 2017, but it droped in 2018.
- Notably, there is a huge increase from 275 to 51509 in year 2014.

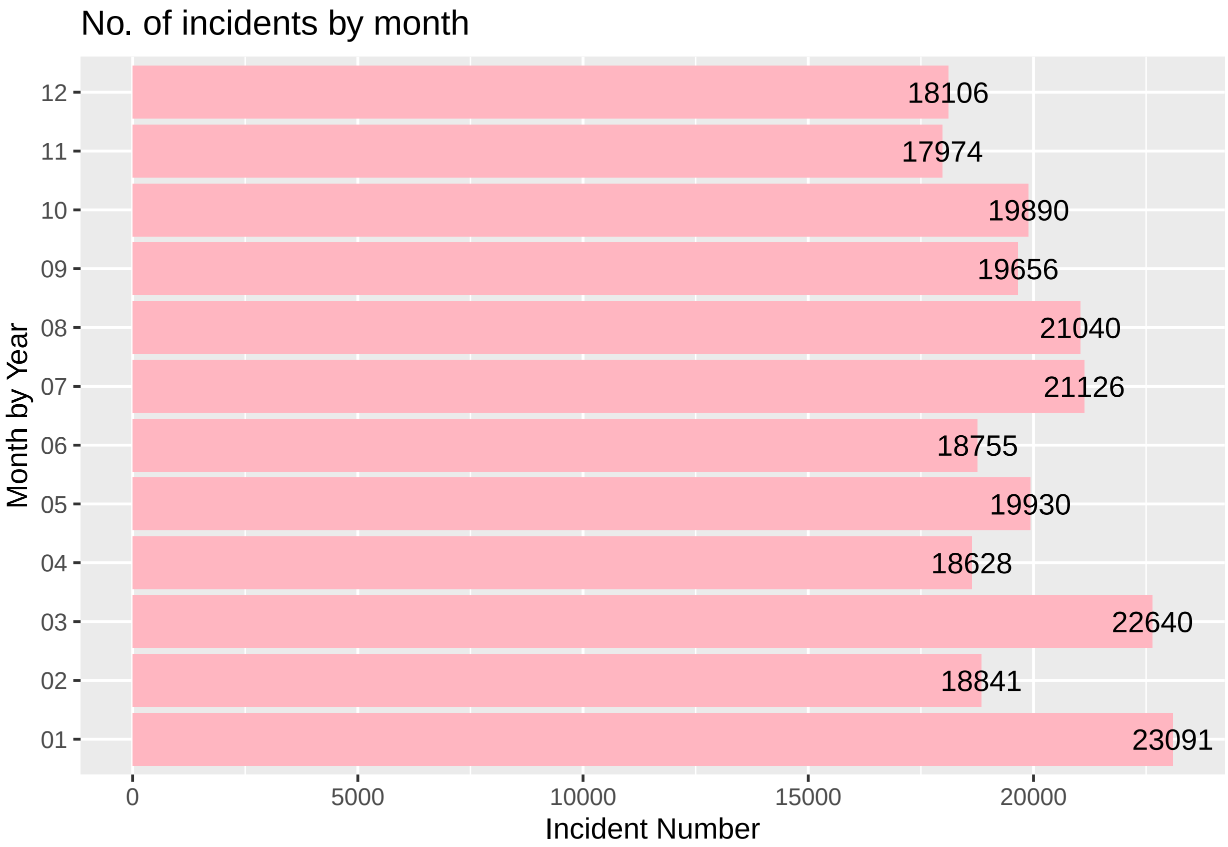
```
incidents_by_year <- gun %>%
  mutate(year = isoyear(date)) %>%
  group_by(year) %>%
  count()
ggplot(incidents_by_year, aes(x= year, y= n)) +
  geom_bar(stat="identity", fill = "light blue") + # change plot color to light blue
  xlab("Year") +
  ylab("Incident Number") +
  ggtitle("No. of Incidents by Year") +
  geom_text(aes(label = n)) # Add the count for each year as an annotation label.
```



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.

- using the whole data would make the graph nasty, so i use monthly data that exceeds 5000 incidents

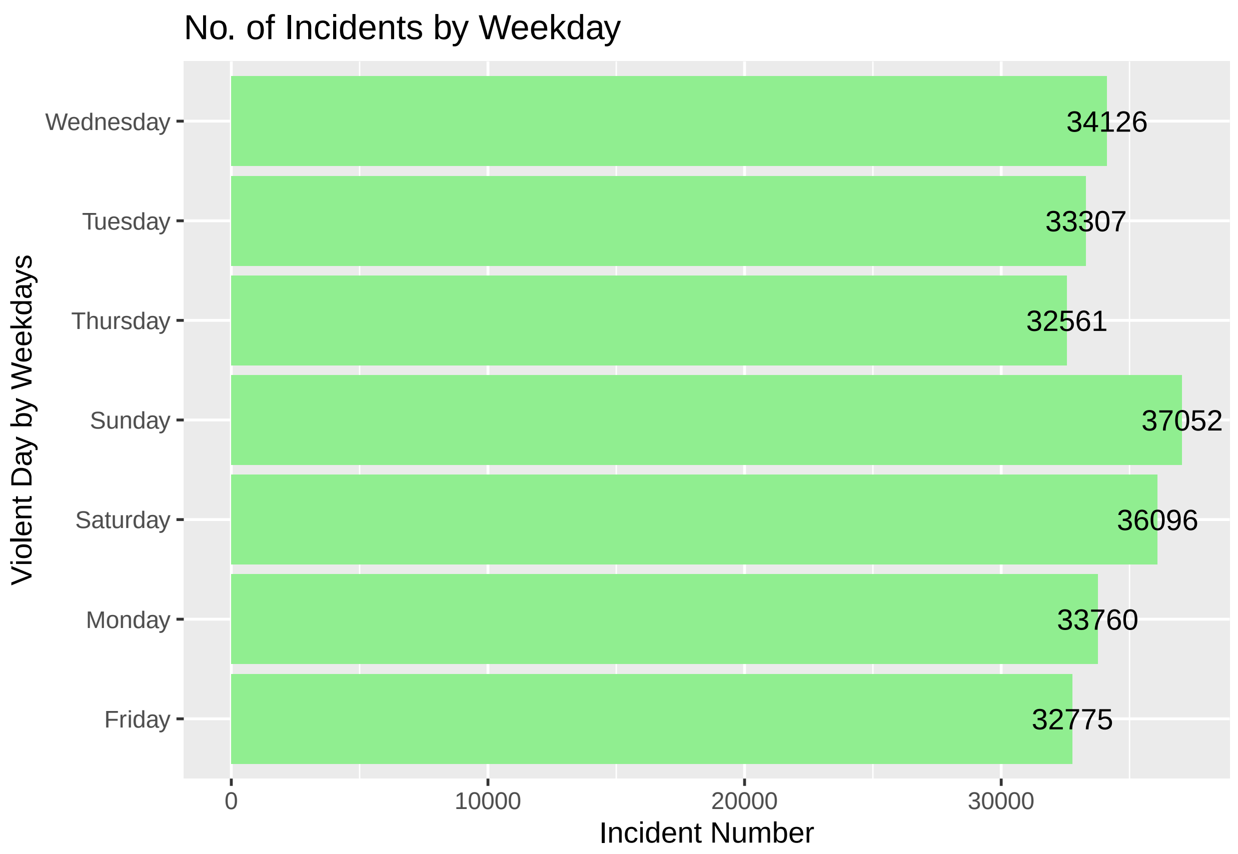
```
mytable <- table(format(gun$date,"%m")) # count the number of rows for each month
# cited from https://stackoverflow.com/questions/19737257/count-frequency-of-observations-by-month
incidents_by_month <- as.data.frame(mytable) # tranform table into a data frame
incidents_by_month <- incidents_by_month %>%
  rename(month = Var1, n = Freq)
ggplot(incidents_by_month, aes(x= month, y= n)) +
  geom_bar(stat="identity", fill = "light pink") + # change plot color to light blue
  xlab("Month by Year") +
  ylab("Incident Number") +
  coord_flip() +
  ggtitle("No. of incidents by month") +
  geom_text(aes(label = n)) # Add the count for each year as an annotation label.
```



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

- According to the graph, Sunday is the weekday with most incidents happended.

```
gun$weekday = weekdays(gun$date)
incidents_by_weekday <- gun %>%
  group_by(weekday) %>%
  count()
ggplot(incidents_by_weekday, aes(x= weekday, y= n)) +
  geom_bar(stat="identity", fill = "light green") + # change plot color to light blue
  xlab("Violent Day by Weekdays") +
  ylab("Incident Number") +
  ggtitle("No. of Incidents by Weekday") +
  coord_flip() +
  geom_text(aes(label = n)) # Add the count for each year as an annotation label.
```



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

- according to the graphs above, i found that year 2014 to year 2017 have incidents over 50000, January is the most violent month, and Sunday is the most violent weekday.

4 3

4.1 Characteristics of Gun violence incidents

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

- 1.372442 guns involved in an incident on average.

```
gun %>%
  select(n_guns_involved) %>%
  summarise(mean_guns_involved = mean(n_guns_involved, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   mean_guns_involved
##                <dbl>
## 1               1.37
```

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

```
# remove na and unknown
gv_1<-gun %>%
  filter(!is.na(date))%>%
  mutate(date =ymd(date))
```

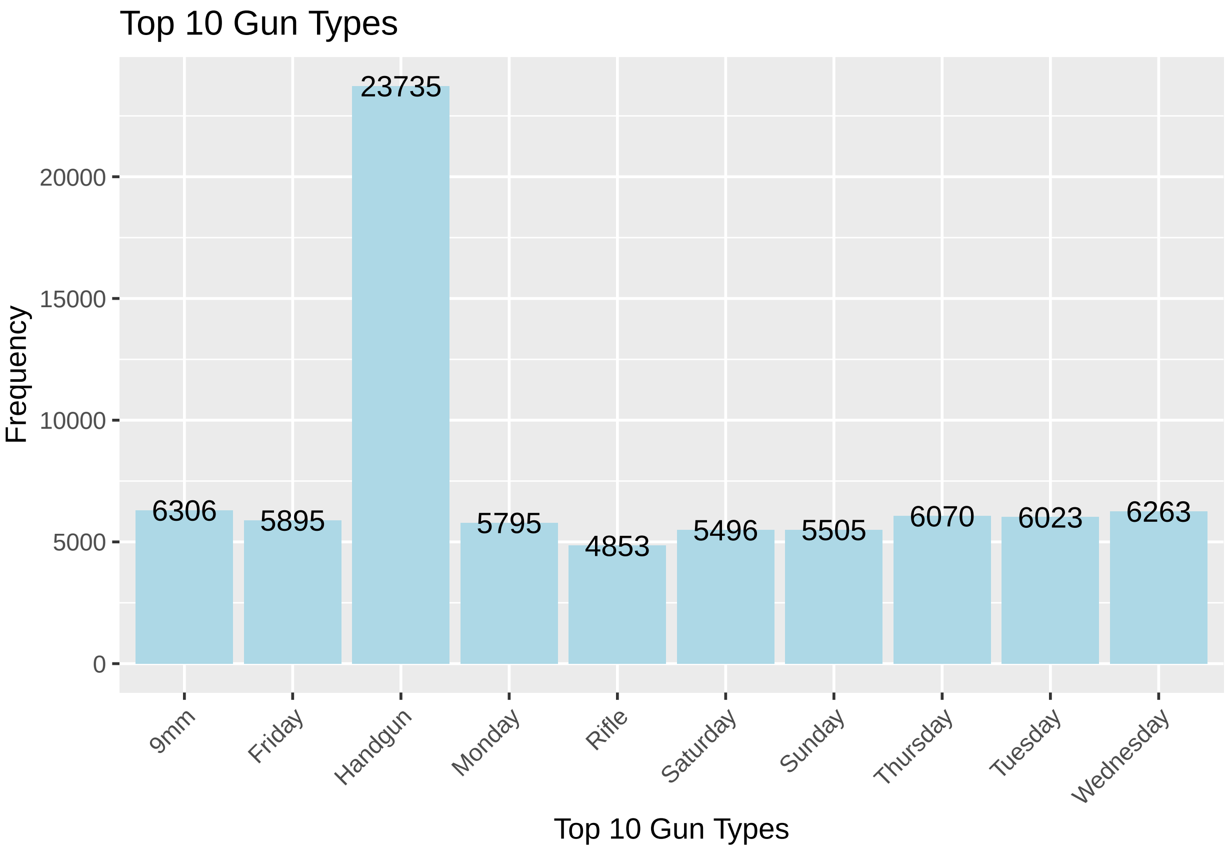
```
gv_remove_unknown<-gv_1 %>%
  filter(!is.na(gun_type)) %>%
  filter(gun_type!="")%>%
  filter(!str_detect(gun_type, "Unknown")) # filter out nas and unknown values

# split values in a sting
gv_gun_type<-gv_remove_unknown %>%
        mutate(gun_type = strsplit(gun_type, "\\|\\|")) %>%
        unnest(gun_type) %>%
        group_by(incident_id) %>%
        mutate(row = row_number()) %>%
        spread(row, gun_type) # split sting values with multiple descriptions

#gv_gun_type %>%
 #   mutate(length=str_length("1")) %>%
 # arrange(-length)
# remove characters in string
gv_gun_type <- gv_gun_type[, c(29:110)]
gv_gun_type <- apply(gv_gun_type,2,function(x) gsub("[0-9]+::","",x))
gv_gun_type <- apply(gv_gun_type,2,function(x) gsub("0:","",x))
# cited from https://stackoverflow.com/questions/48953295/replace-a-specific-strings-from-multiple-colu

# filter out top 10 gun types involved
gun_type_occurences<-table(unlist(gv_gun_type))
gun_type_occurences <- as.data.frame(gun_type_occurences)
top_10_gun_type_occurences <- gun_type_occurences%>%
  arrange(desc(Freq))%>%
  head(10)

# plotting
ggplot(top_10_gun_type_occurences, aes(x= Var1, y= Freq)) +
  geom_bar(stat="identity", fill = "light blue") + # change plot color to light blue
  xlab("Top 10 Gun Types") +
  ylab("Frequency") +
  geom_text(aes(label = Freq)) +
  ggtitle("Top 10 Gun Types") +
  theme(axis.text.x=element_text(angle=45, hjust=1)) #rotate x axis for better presentation
```



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

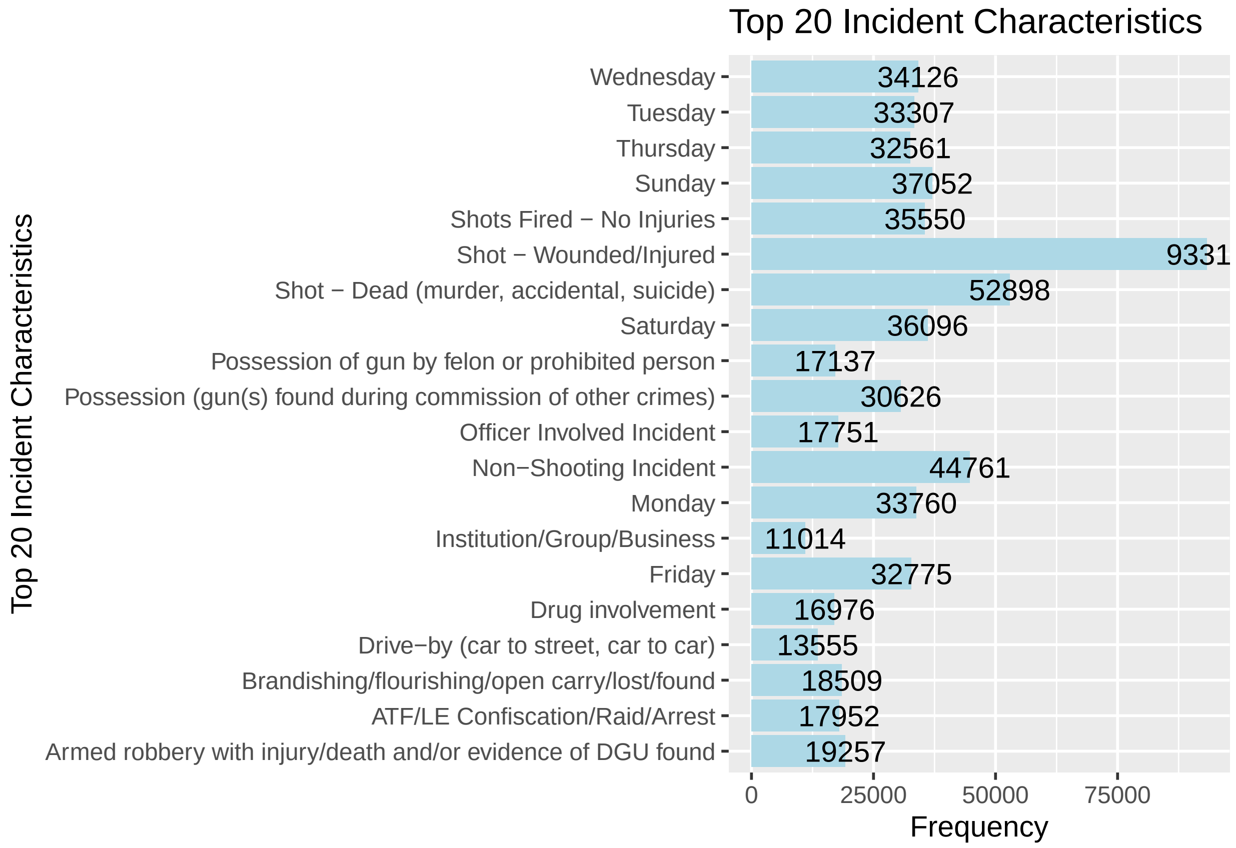
- this variable describes the means of committing the crime, and the consequence of the crime
- there are multiple characteristics for each incident because the suspects might shot multiple times and resulted in different consequence of the victims

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)

```
# split values in a string
incident_char <- gun %>%
        mutate(incident_characteristics = strsplit(incident_characteristics, "\\|\\|")) %>%
        unnest(incident_characteristics) %>%
        group_by(incident_id) %>%
        mutate(row = row_number()) %>%
        spread(row, incident_characteristics)

# filter out top 20 incident characteristics
incident_char <- incident_char[, c(29:47)]
incident_char_occurences<-table(unlist(incident_char))
incident_char_occurences <- as.data.frame(incident_char_occurences)
top_20_incident_char_occurences <- incident_char_occurences%>%
  arrange(desc(Freq))%>%
  head(20)

# plotting
ggplot(top_20_incident_char_occurences, aes(x= Var1, y= Freq)) +
  geom_bar(stat="identity", fill = "light blue") + # change plot color to light blue
  xlab("Top 20 Incident Characteristics") +
  ylab("Frequency") +
  ggtitle("Top 20 Incident Characteristics") +
  geom_text(aes(label = Freq)) +
  coord_flip()
```



5 4

5.1 Suspects Characteristics

1. Explore the following variables and write a short paragraph of what they mean and how they are connect 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.

- these four variables describe the demographic information of both the suspects and victims involved in each incident. They also include information about the consequence of the shooting.
- they are difficult to interpret because they mix the information of criminals and victims together.

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

```
# split string with multiple patterns
participant_type <- gun %>%
     mutate(participant_type = strsplit(participant_type, "[|||]+"))%>%
     unnest(participant_type) %>%
     group_by(incident_id) %>%
     mutate(row = row_number()) %>%
     spread(row, participant_type)
```

```
# remove characters in string
participant_type <- participant_type[, c(29:131)]
participant_type <- apply(participant_type,2,function(x) gsub("[0-9]+::","",x))
participant_type <- apply(participant_type,2,function(x) gsub("[0-9]+:","",x))

# build a table
participant_type_table <-table(unlist(participant_type))
participant_type_table <- as.data.frame(participant_type_table)
participant_type_table
```

```
##              Var1   Freq
## 1          Friday  32775
## 2          Monday  33760
## 3        Saturday  36096
## 4 Subject-Suspect 199262
## 5          Sunday  37052
## 6        Thursday  32561
## 7         Tuesday  33307
## 8          Victim 193060
## 9       Wednesday  34126
```

```
# average no. of suspects per incident =
199262/nrow(gun)
```

```
## [1] 0.8313772
```

```
# average no. of victims per incident =
193060/nrow(gun)
```

```
## [1] 0.8055007
```

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

```
# unnest participant_type
participant_type_unnest <- gun %>%
        mutate(participant_type = strsplit(participant_type, "[|||]+")
               )%>%
        unnest(participant_type)
# unnest suspect
suspect_unnest <- participant_type_unnest %>%
    filter( str_detect(participant_type ,"Suspect")  )%>%
     select(incident_id, participant_age, participant_gender,participant_type)
# unnest suspect age
suspect_age_unnest <- suspect_unnest %>%
  mutate(participant_age = strsplit(participant_age, "[|||]+")) %>%
        unnest(participant_age)
# unnest suspect gender
suspect_gender_unnest <- suspect_age_unnest %>%
  mutate(age_2 = str_extract( participant_age ,"(.*)::"),
         type_2 =  str_extract( participant_type ,"(.*)::"),
         gender_2 = str_extract( participant_gender ,"(.*)::")
        )
# suspect data frame
suspect_final <- suspect_gender_unnest %>%
  filter(age_2 == type_2 & type_2 == gender_2)
```

```
suspect_final2 <- suspect_final  %>%
  select(incident_id, participant_age, participant_gender)
head(suspect_final2)
```

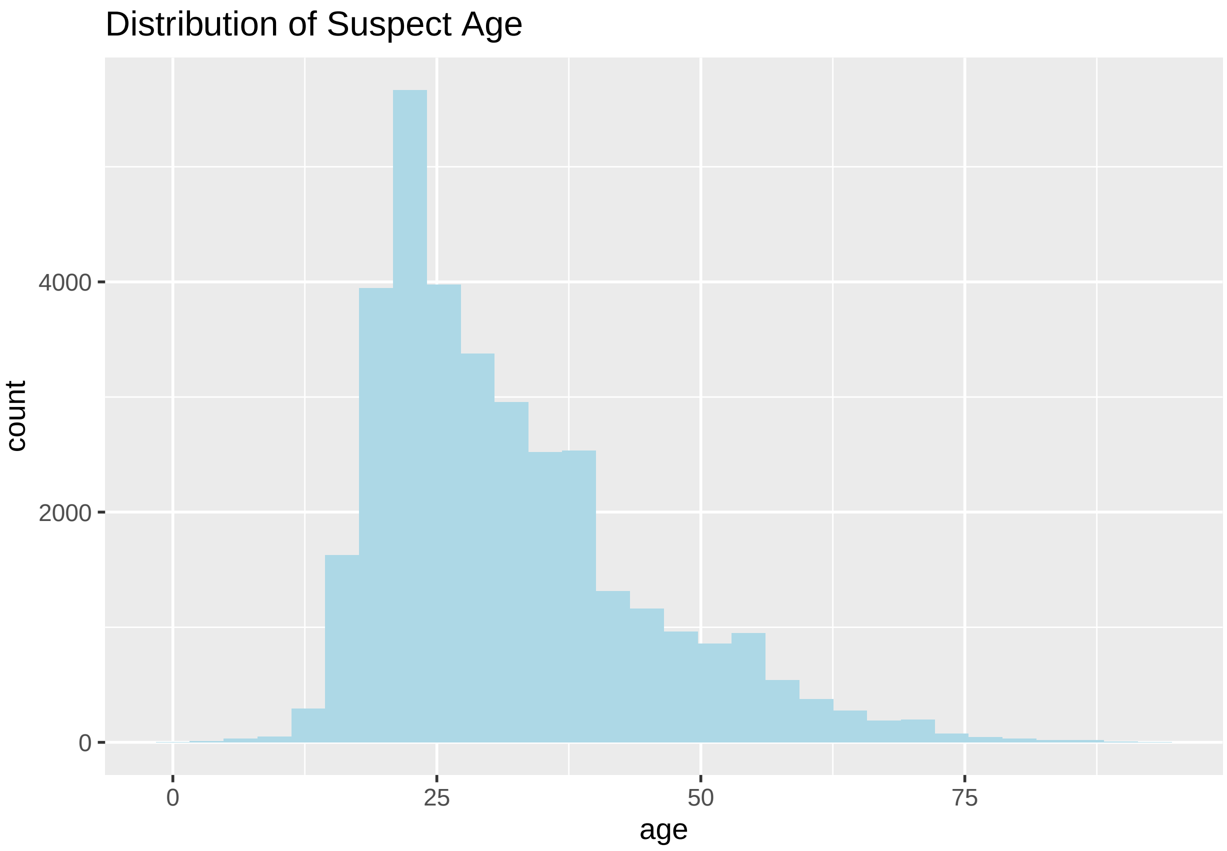
```
## # A tibble: 6 x 3
##   incident_id participant_age participant_gender
##         <dbl> <chr>           <chr>
## 1      893251 0::22           0::Male
## 2      490395 4::22           4::Female
## 3      479821 4::35           4::Male
## 4      964582 1::24           1::Male
## 5       96019 0::35           0::Female
## 6       92506 0::22           0::Male
```

4. Show the distribution of suspects age, crop your plot if you find any suspect over the age of 100

```
suspect_final2 <- suspect_final2  %>%
  mutate(
  age = as.numeric( str_remove( str_extract( participant_age ,"::(.*)"),"::"))
  )
# cited from https://stringr.tidyverse.org/reference/str_remove.html

#plot the distribution of age
ggplot(suspect_final2, aes(age)) +
  geom_histogram(fill = "light blue") +
  ggtitle("Distribution of Suspect Age")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
#no age over 100
suspect_final2 %>% filter(age > 100)
```

```
## # A tibble: 0 x 4
## # ... with 4 variables: incident_id <dbl>, participant_age <chr>,
## #   participant_gender <chr>, age <dbl>
```

6. What percentage of suspects are male (exclude missing values)?

- 95.37% suspects are male.

```
suspect_final2 <- suspect_final2  %>%
  mutate(
    gender =  str_remove( str_extract( participant_gender ,"::(.*)"),"::")
)

suspect_final2 %>% summarise(pct = mean(gender == "Male") * 100)
```

```
## # A tibble: 1 x 1
##     pct
##   <dbl>
## 1  95.4
```

6. How many different status are there?

- there are 12 status in total.

```
participant_status_unnest <- gun %>%
        mutate(participant_status = strsplit(participant_status, "[|||]+")
               )%>%
        unnest(participant_status)

participant_status_unnest <- participant_status_unnest %>%
  mutate(
  status =  str_remove( str_extract( participant_status ,"::(.*)"),"::")
)

participant_status_unnest %>%
  group_by(status) %>%
  count() %>%
  na.omit()
```

```
## # A tibble: 12 x 2
## # Groups:   status [12]
##    status                           n
##    <chr>                        <int>
##   1 Arrested                    10169
##   2 Injured                    113332
##   3 Injured, Arrested            3467
##   4 Injured, Unharmed              31
##   5 Injured, Unharmed, Arrested    22
##   6 Killed                      59386
##   7 Killed, Arrested               51
##   8 Killed, Injured                10
##   9 Killed, Unharmed               21
## 10 Killed, Unharmed, Arrested     14
## 11 Unharmed                    100822
## 12 Unharmed, Arrested           85388
```

7. What percentage of all suspects got arrested? Be careful for some suspects there are more than 1 categories.

- 51.7% of the suspects got arrested.

```
participant_type_suspect <- participant_type_unnest %>%
  filter( str_detect(  participant_type ,"Suspect")  )%>%
     select(incident_id,participant_type, participant_status)

participant_status_suspect <- participant_type_suspect %>%
  mutate(participant_status = strsplit(participant_status, "[|||]+")) %>%
        unnest(participant_status)

participant_status_suspect2 <- participant_status_suspect %>%
  mutate(status_2  =  str_extract( participant_status ,"(.*)::"),
                      type_2 =  str_extract( participant_type ,"(.*)::"),
           status =  str_remove( str_extract( participant_status ,"::(.*)"),"::") )

suspect_status_final <- participant_status_suspect2 %>%
                        filter(type_2 == status_2 )

suspect_status_final <- suspect_status_final %>%
                        mutate(arrested = str_detect(status, "Arrested"))

suspect_status_final %>% summarise(pct = mean(arrested) *100)
```

```
## # A tibble: 1 x 1
##     pct
##   <dbl>
## 1  51.7
```

6 5

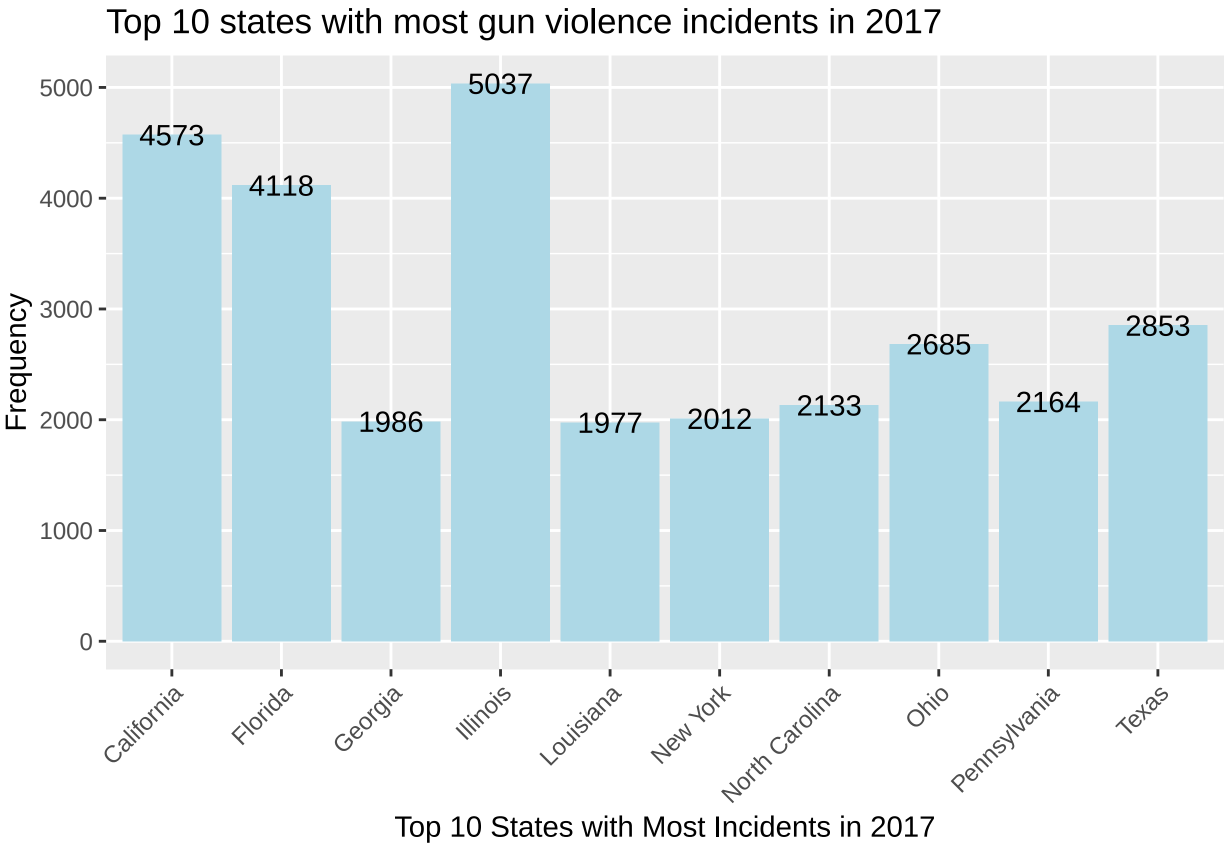
6.1 Geographical Variation

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

- illinois.

```
incidents_by_year <- gun %>%
  mutate(year = isoyear(date)) %>%
  group_by(year) %>%
  count()
incident2017 <- gun %>%
  mutate(year = isoyear(date)) %>%
  filter(year == 2017)  %>%
  group_by(state) %>%
  count() %>%
  arrange(desc(n)) %>%
  head(10)

# plotting
ggplot(incident2017, aes(x= state, y= n)) +
  geom_bar(stat="identity", fill = "light blue") + # change plot color to light blue
  xlab("Top 10 States with Most Incidents in 2017") +
  ylab("Frequency") +
  ggtitle("Top 10 states with most gun violence incidents in 2017") +
  geom_text(aes(label = n)) +
  theme(axis.text.x=element_text(angle=45, hjust=1)) #rotate x axis for better presentation
```



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

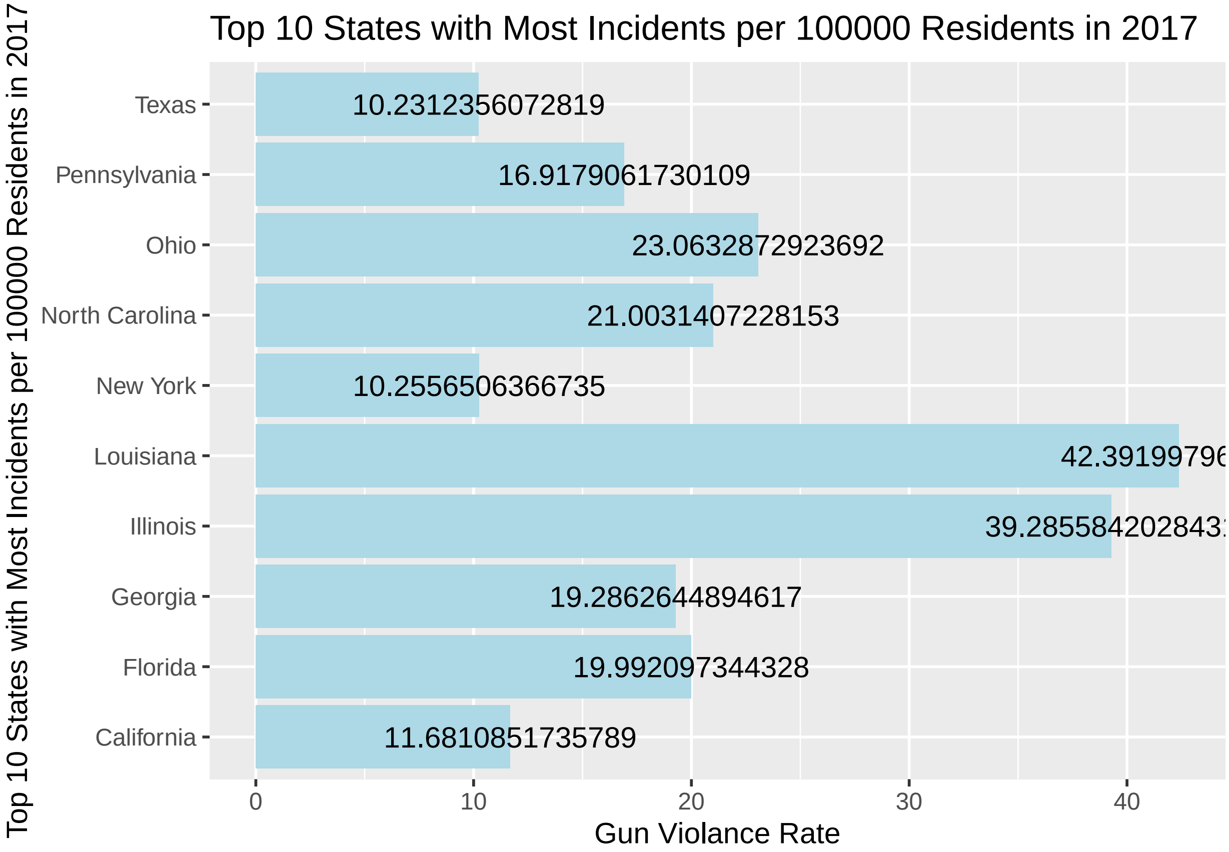
```
# get population data
library(tidycensus)
ACS_VARS_18<-load_variables(2018,"acs5",cache=TRUE)
acs_data<-get_acs(
  geography = "state",
  variables=c("B01003_001"),
  year=2018
)
```

```
## Getting data from the 2014-2018 5-year ACS
```

```
acs_data <- acs_data %>%
  rename(state=NAME,population = estimate)

# combining two data frames and calculate incidents by a 100,000 inhabitants
incident2017_by_population <- merge(incident2017, acs_data, by = "state")
incident2017_by_population <- incident2017_by_population %>%
  mutate(incident_adjusted = (n/population)*100000) %>%
  arrange(desc(incident_adjusted))
top_10_incident2017_by_population <- incident2017_by_population %>%
  head(10)

#plotting
ggplot(top_10_incident2017_by_population,  aes(x= state, y= incident_adjusted)) +
  geom_bar(stat="identity", fill = "light blue") + # change plot color to light blue
  xlab("Top 10 States with Most Incidents per 100000 Residents in 2017") +
  ylab("Gun Violance Rate") +
  ggtitle("Top 10 States with Most Incidents per 100000 Residents in 2017") +
  geom_text(aes(label = incident_adjusted)) +
  coord_flip()
```



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

```
# getting us map
states <- st_as_sf(map("state", plot = FALSE, fill = TRUE))
sf_use_s2(FALSE)
```

```
## Spherical geometry (s2) switched off
```

```
states_centroids <- st_coordinates(st_centroid(states))
```

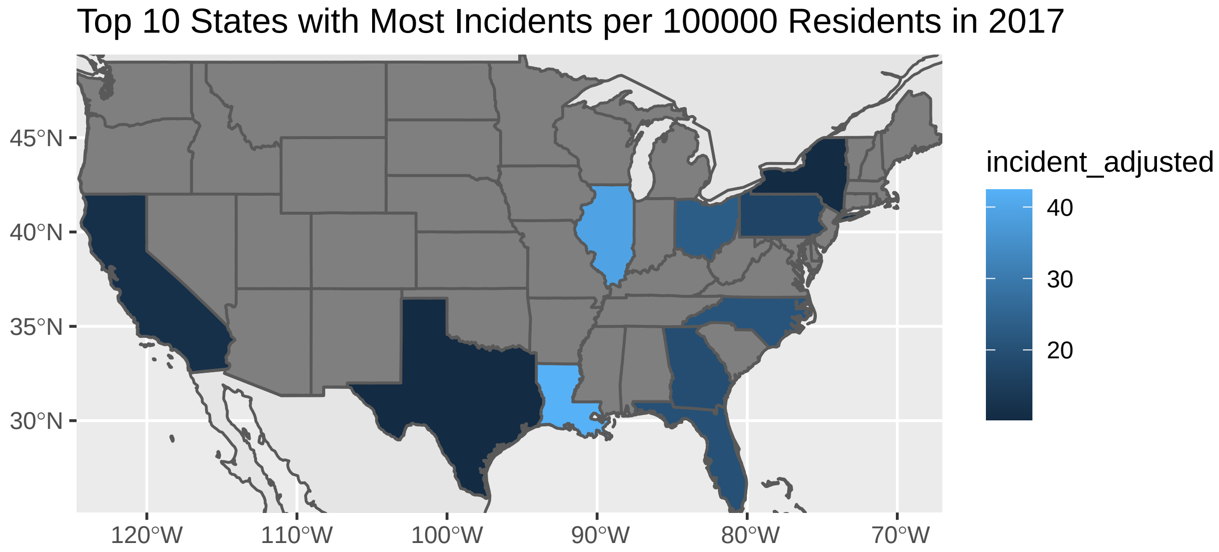
```
## Warning in st_centroid.sf(states): st_centroid assumes attributes are constant
## over geometries of x
```

```
## Warning in st_centroid.sfc(st_geometry(x), of_largest_polygon =
## of_largest_polygon): st_centroid does not give correct centroids for longitude/
## latitude data
```

```r
world <- ne_countries(scale = "medium",
returnclass = "sf")
states<- states %>%
        mutate(X = states_centroids[,1],
               Y = states_centroids[,2])
# matching state with incidents numbers.
top_10_incident2017_by_population$ID <- tolower(top_10_incident2017_by_population$state)
states2 <- states %>% left_join(top_10_incident2017_by_population)
```

```
## Joining, by = "ID"
```

```r
# plotting
ggplot(data = world) +
  geom_sf() +
  geom_sf(data = states2,aes( fill = incident_adjusted) )+
  coord_sf(xlim = c(-124.6813,-67.00742),
  ylim = c(25.12993,49.38323), expand = FALSE) +
  ggtitle("Top 10 States with Most Incidents per 100000 Residents in 2017")
```



6.2 Mass Shooting

1. Create a new data frame of mass shootings

```r
incident_characteristics_unnest <- gun %>%
  mutate(incident_characteristics = strsplit(incident_characteristics, "[|||]+"))%>%
        unnest(incident_characteristics)

mass_shooting <- incident_characteristics_unnest %>%
  filter( str_detect( incident_characteristics ,"Mass Shooting")  )
```

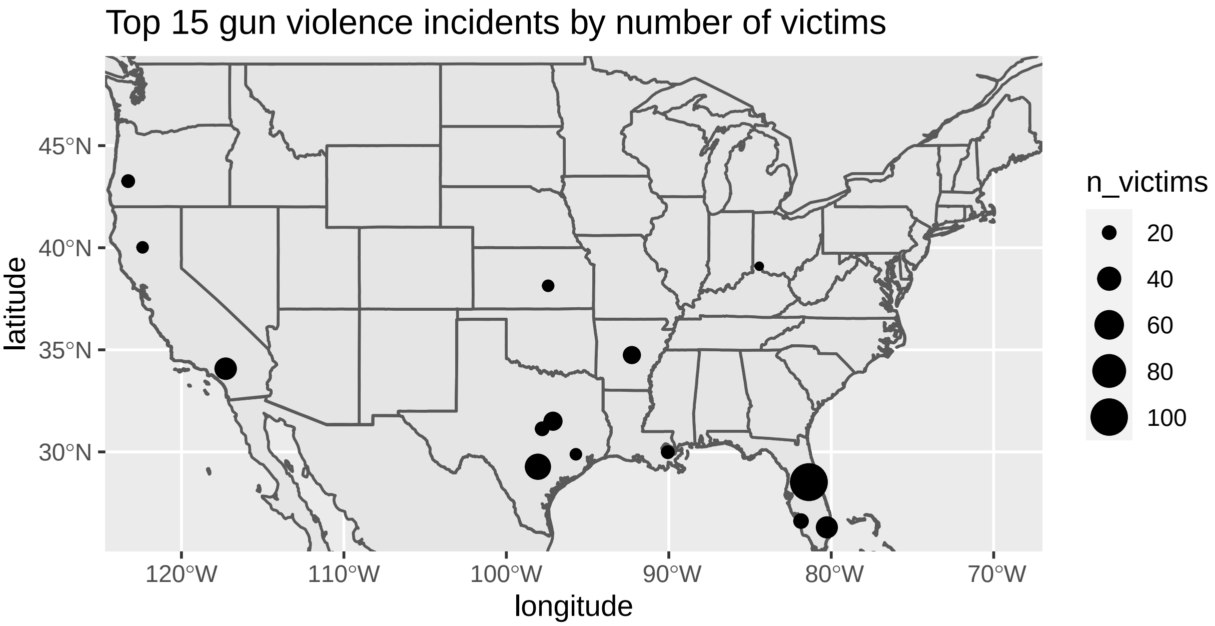
2. Show the top 15 incidents by number of victims in a map as points

```r
# matching the top 15 incidents with states
mass_shooting_victims <- mass_shooting %>%
  mutate(n_victims =  n_killed + n_injured) %>%
  arrange(desc(n_victims))
top_15_mass_shooting_victims <- head(mass_shooting_victims, 15)
top_15_mass_shooting_victims$ID <- tolower(top_15_mass_shooting_victims$state)
```

```
states3 <- states %>% left_join(top_15_mass_shooting_victims)
```

```
## Joining, by = "ID"
```

```
# plotting
ggplot(data = world) +
  geom_sf() +
  geom_sf(data = states3 )+
  geom_point(data = top_15_mass_shooting_victims,
  aes(longitude ,latitude, size = n_victims), fill = "red") +
  coord_sf(xlim = c(-124.6813,-67.00742),
  ylim = c(25.12993,49.38323), expand = FALSE) +
  ggtitle("Top 15 gun violence incidents by number of victims")
```



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

- we can find that in mass shootings, the average age is lower, the male percentage is higher, the arrested percentage is higher and there is more number of shooters.
- However, the average percentage of killed number is lower.

```
participant_type_unnest <- gun %>%
        mutate(participant_type = strsplit(participant_type, "[|||]+")
               )%>%
        unnest(participant_type)

suspect <- participant_type_unnest %>%
  filter( str_detect(  participant_type ,"Suspect")  )
incident_characteristics_unnest <- suspect %>%
  mutate(incident_characteristics = strsplit(incident_characteristics, "[|||]+")) %>%
        unnest(incident_characteristics)
incident_characteristics_mass_shooting <- incident_characteristics_unnest %>%
  mutate(Mass_shotting= str_detect( incident_characteristics ,"Mass Shooting")  )

mass_shooting_participant_age <- incident_characteristics_mass_shooting %>%
    mutate(participant_age = strsplit(participant_age, "[|||]+"))%>%
    unnest(participant_age)

mass_shooting_participant_age <- mass_shooting_participant_age %>%
  mutate(age_2 = str_extract( participant_age ,"(.*)::"),
         type_2 =  str_extract( participant_type ,"(.*)::"))

mass_shooting_participant_age <- mass_shooting_participant_age %>%
  filter(age_2 == type_2)

mass_shooting_char <- mass_shooting_participant_age %>%
  select( incident_id, n_killed,n_injured,Mass_shotting,type_2,participant_status,
                  participant_gender, participant_age,participant_type )

mass_shooting_char <- mass_shooting_char %>%
  mutate(participant_gender = strsplit(participant_gender, "[|||]+"))%>%
  unnest(participant_gender) %>%
  mutate(gender_2 = str_extract( participant_gender ,"(.*)::")) %>%
  filter(gender_2 == type_2)


mass_status_unnest <- mass_shooting_char %>%
  mutate(participant_status = strsplit(participant_status, "[|||]+")) %>%
  unnest(participant_status)

mass_status_unnest <- mass_status_unnest %>%
  mutate(status_2 = str_extract( participant_status ,"(.*)::")) %>%
  filter(status_2 == type_2)

final_1 <- mass_status_unnest %>%
  mutate(
  age = as.numeric( str_remove( str_extract( participant_age ,"::(.*)"),"::")),
  gender =  str_remove( str_extract( participant_gender ,"::(.*)"),"::"),
  status =  str_remove( str_extract( participant_status,"::(.*)"),"::"),
 pct_killed = n_killed /(n_killed + n_injured) *100
  )%>%
  mutate(
  status_arrest = str_detect(status , "Arrested")
  ) %>%
  select( incident_id, age, gender, status_arrest,  pct_killed ,Mass_shotting)


final_2 <- mass_status_unnest %>% group_by(incident_id) %>%summarise(nshooters = length(unique(particip

final <- final_1 %>% left_join(final_2)
```

```
## Joining, by = "incident_id"
```

```
final<-
final %>%
  group_by(Mass_shotting) %>%
  summarise(
    age = mean(age),
    male_pct = mean(gender == "Male")*100,
    arrest_pct = mean(status_arrest) *100,
     pct_killed = mean( pct_killed, na.rm = T),
    n_shooters = mean( nshooters)) %>%
  drop_na()

final
```

```
## # A tibble: 2 x 6
##   Mass_shotting   age male_pct arrest_pct pct_killed n_shooters
##   <lgl>         <dbl>    <dbl>      <dbl>      <dbl>      <dbl>
## 1 FALSE          29.5     92.2       74.9       49.0       1.88
## 2 TRUE           26.3     96.6       75.8       30.0       2.33
```