[GUN VIOLENCE] Incidents\_by\_state

1585107 이지인

2018년 12월 3일

( df1 <- df1 %>%  
 rename(long=longitude, lat=latitude) ) # long, lat 변수 통일

## # A tibble: 173,739 x 15  
## incident\_id year month day weekdays state city\_or\_county n\_killed  
## <int> <int> <int> <int> <chr> <chr> <chr> <int>  
## 1 176097 2016 8 14 Sun New ~ New York (Man~ 2  
## 2 271939 2015 1 1 Thurs Ohio Cleveland 0  
## 3 271945 2015 1 1 Thurs Illi~ Chicago 0  
## 4 271947 2015 1 1 Thurs Illi~ Chicago 0  
## 5 271949 2015 1 1 Thurs Illi~ Chicago 0  
## 6 271951 2015 1 1 Thurs Neva~ Las Vegas 1  
## 7 271954 2015 1 1 Thurs Neva~ Las Vegas 0  
## 8 271970 2015 1 1 Thurs Cali~ Downey 0  
## 9 271975 2015 1 1 Thurs Cali~ San Jose 0  
## 10 271979 2015 1 1 Thurs Cali~ Grand Terrace 1  
## # ... with 173,729 more rows, and 7 more variables: n\_injured <int>,  
## # congressional\_district <int>, lat <dbl>, long <dbl>,  
## # n\_guns\_involved <int>, state\_house\_district <int>,  
## # state\_senate\_district <int>

( a <- df1 %>%   
 count(state) %>%   
 rename(acc=n) ) # acc 주별(51개) 사건 수 변수

## # A tibble: 51 x 2  
## state acc  
## <chr> <int>  
## 1 Alabama 3804  
## 2 Alaska 1142  
## 3 Arizona 1621  
## 4 Arkansas 2077  
## 5 California 11439  
## 6 Colorado 2409  
## 7 Connecticut 2320  
## 8 Delaware 1083  
## 9 District of Columbia 2177  
## 10 Florida 11059  
## # ... with 41 more rows

code <- code[,c(1,2,4,5,6)]  
head(code) # state code lat long pop(인구)

## # A tibble: 6 x 5  
## state code lat long pop  
## <chr> <int> <dbl> <dbl> <int>  
## 1 Alabama 1 32.8 -86.8 4874747  
## 2 Alaska 2 61.4 -152. 739795  
## 3 Arizona 4 33.7 -111. 7016270  
## 4 Arkansas 5 35.0 -92.4 3004279  
## 5 California 6 36.1 -120. 39536653  
## 6 Colorado 8 39.1 -105. 5607154

( b <- full\_join(code,a) )

## Joining, by = "state"

## # A tibble: 51 x 6  
## state code lat long pop acc  
## <chr> <int> <dbl> <dbl> <int> <int>  
## 1 Alabama 1 32.8 -86.8 4874747 3804  
## 2 Alaska 2 61.4 -152. 739795 1142  
## 3 Arizona 4 33.7 -111. 7016270 1621  
## 4 Arkansas 5 35.0 -92.4 3004279 2077  
## 5 California 6 36.1 -120. 39536653 11439  
## 6 Colorado 8 39.1 -105. 5607154 2409  
## 7 Connecticut 9 41.6 -72.8 3588184 2320  
## 8 Delaware 10 39.3 -75.5 961939 1083  
## 9 District of Columbia 11 38.9 -77.0 693972 2177  
## 10 Florida 12 27.8 -81.7 20984400 11059  
## # ... with 41 more rows

# 인구 수를 고려했을 때 사건 수 - 인구 100000명 당 사고 발생 횟수  
# accidents per 100000: acc\_pop   
( b <- b %>% # b - state code lat long pop acc acc\_pop  
 filter(!is.na(acc), !is.na(pop)) %>%  
 mutate( acc\_pop = round( (acc/pop)\*100000 ) ) )

## # A tibble: 51 x 7  
## state code lat long pop acc acc\_pop  
## <chr> <int> <dbl> <dbl> <int> <int> <dbl>  
## 1 Alabama 1 32.8 -86.8 4874747 3804 78  
## 2 Alaska 2 61.4 -152. 739795 1142 154  
## 3 Arizona 4 33.7 -111. 7016270 1621 23  
## 4 Arkansas 5 35.0 -92.4 3004279 2077 69  
## 5 California 6 36.1 -120. 39536653 11439 29  
## 6 Colorado 8 39.1 -105. 5607154 2409 43  
## 7 Connecticut 9 41.6 -72.8 3588184 2320 65  
## 8 Delaware 10 39.3 -75.5 961939 1083 113  
## 9 District of Columbia 11 38.9 -77.0 693972 2177 314  
## 10 Florida 12 27.8 -81.7 20984400 11059 53  
## # ... with 41 more rows

( usmap1 <- usmap %>%  
 tbl\_df %>%  
 rename(state = region) %>% # state 변수 통일   
 select(long:state) )

## # A tibble: 15,537 x 5  
## long lat group order state   
## \* <dbl> <dbl> <dbl> <int> <chr>   
## 1 -87.5 30.4 1 1 alabama  
## 2 -87.5 30.4 1 2 alabama  
## 3 -87.5 30.4 1 3 alabama  
## 4 -87.5 30.3 1 4 alabama  
## 5 -87.6 30.3 1 5 alabama  
## 6 -87.6 30.3 1 6 alabama  
## 7 -87.6 30.3 1 7 alabama  
## 8 -87.6 30.3 1 8 alabama  
## 9 -87.7 30.3 1 9 alabama  
## 10 -87.8 30.3 1 10 alabama  
## # ... with 15,527 more rows

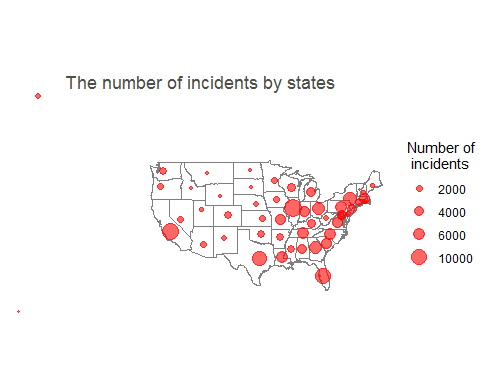
# b1 - state acc acc\_pop  
( b1 <- b %>%  
 mutate( state = tolower(state) ) %>%  
 select( state, acc, acc\_pop ) )

## # A tibble: 51 x 3  
## state acc acc\_pop  
## <chr> <int> <dbl>  
## 1 alabama 3804 78  
## 2 alaska 1142 154  
## 3 arizona 1621 23  
## 4 arkansas 2077 69  
## 5 california 11439 29  
## 6 colorado 2409 43  
## 7 connecticut 2320 65  
## 8 delaware 1083 113  
## 9 district of columbia 2177 314  
## 10 florida 11059 53  
## # ... with 41 more rows

# b2 - usmap + acc acc\_pop   
( b2 <- right\_join(usmap1, b1, by="state") )

## # A tibble: 15,539 x 7  
## long lat group order state acc acc\_pop  
## <dbl> <dbl> <dbl> <int> <chr> <int> <dbl>  
## 1 -87.5 30.4 1 1 alabama 3804 78  
## 2 -87.5 30.4 1 2 alabama 3804 78  
## 3 -87.5 30.4 1 3 alabama 3804 78  
## 4 -87.5 30.3 1 4 alabama 3804 78  
## 5 -87.6 30.3 1 5 alabama 3804 78  
## 6 -87.6 30.3 1 6 alabama 3804 78  
## 7 -87.6 30.3 1 7 alabama 3804 78  
## 8 -87.6 30.3 1 8 alabama 3804 78  
## 9 -87.7 30.3 1 9 alabama 3804 78  
## 10 -87.8 30.3 1 10 alabama 3804 78  
## # ... with 15,529 more rows

# 어느 주에서 (인구대비) 총기사고가 많이 발생했는가  
# figure 1.1) Incidents by states 주별 사고 수 지도 - size옵션  
ggplot(b2, aes(long, lat))+   
 geom\_polygon( aes(group=group), data=usmap,   
 fill=NA, color="grey50")+  
 geom\_point( aes(long,lat, size=acc),   
 data=b, colour="red", alpha=0.6)+  
 scale\_size\_area(breaks=c(2000,4000,6000,10000),  
 name="Number of \n incidents")+  
 labs(title = "The number of incidents by states")+  
 coord\_quickmap()+  
 theme\_void()+  
 theme( plot.title = element\_text(hjust=0.5, vjust=-1,  
 colour="#4e4d47") )



# 주별 사고수 편차가 크기 때문에 scale조정이 필요.  
summary(b2$acc)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 222 2320 4156 4889 6469 13600

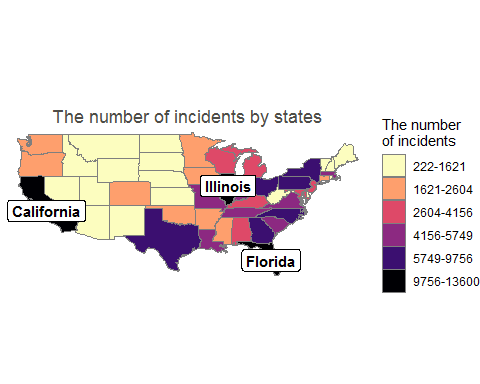
# scale 조정  
quantiles <- quantile(b2$acc,  
 probs=seq(0, 1,length.out = 6+1))  
labels<- c()   
for(idx in 1:length(quantiles)){  
 labels <- c(labels, paste0(round(quantiles[idx], 2),  
 "-",  
 round(quantiles[idx+1], 2)))  
}  
labels <- labels[1:length(labels)-1] # 사고수 6개 범주로 나눔  
b2$acc\_quantiles <- cut(b2$acc,  
 breaks = quantiles,  
 labels = labels,  
 include.lowest = T)  
# 총기사고 수 많은 상위 3개 주  
b %>% arrange(desc(acc)) %>% select(state, pop, acc, acc\_pop)

## # A tibble: 51 x 4  
## state pop acc acc\_pop  
## <chr> <int> <int> <dbl>  
## 1 Illinois 12802023 13600 106  
## 2 California 39536653 11439 29  
## 3 Florida 20984400 11059 53  
## 4 Texas 28304596 9756 34  
## 5 New York 19849399 7456 38  
## 6 Ohio 11658609 7275 62  
## 7 Georgia 10429379 6469 62  
## 8 Pennsylvania 12805537 6146 48  
## 9 North Carolina 10273419 6098 59  
## 10 Louisiana 4684333 5749 123  
## # ... with 41 more rows

( text<- b %>%   
 filter(state %in% c("Illinois","California","Florida")) %>%   
 select(state, lat, long) )

## # A tibble: 3 x 3  
## state lat long  
## <chr> <dbl> <dbl>  
## 1 California 36.1 -120.   
## 2 Florida 27.8 -81.7  
## 3 Illinois 40.3 -89.0

# figure 1.2) Incidents by states 주별 사고 수 지도 - fill옵션   
b2 %>%  
 ggplot(aes(long, lat))+  
 geom\_polygon(aes(fill=acc\_quantiles, group=group),  
 colour="grey50", size=0.1)+  
 geom\_label(aes(long, lat, label=state), data=text,  
 fontface = "bold")+  
 coord\_equal()+  
 theme\_void()+  
 labs(title = "The number of incidents by states")+  
 scale\_fill\_viridis( option = "magma",  
 name = "The number\nof incidents ",  
 discrete = TRUE,  
 direction = -1)+  
 theme( plot.title = element\_text(hjust=0.5, colour="#4e4d47" ))



# 인구 수가 많을수록 총기 사고가 많을 수 있다.  
# 주별 인구대비 사고 수 데이터 활용해보자.  
head(b2)

## # A tibble: 6 x 8  
## long lat group order state acc acc\_pop acc\_quantiles  
## <dbl> <dbl> <dbl> <int> <chr> <int> <dbl> <fct>   
## 1 -87.5 30.4 1 1 alabama 3804 78 2604-4156   
## 2 -87.5 30.4 1 2 alabama 3804 78 2604-4156   
## 3 -87.5 30.4 1 3 alabama 3804 78 2604-4156   
## 4 -87.5 30.3 1 4 alabama 3804 78 2604-4156   
## 5 -87.6 30.3 1 5 alabama 3804 78 2604-4156   
## 6 -87.6 30.3 1 6 alabama 3804 78 2604-4156

summary(b2$acc\_pop) # 주별 인구대비 사고 수 편차가 큼

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 40.00 55.00 58.11 67.00 314.00

# 인구 고려시, 사고 수가 많은 주가 달라진다.   
b %>% arrange(desc(acc)) %>% select(state, pop, acc, acc\_pop)

## # A tibble: 51 x 4  
## state pop acc acc\_pop  
## <chr> <int> <int> <dbl>  
## 1 Illinois 12802023 13600 106  
## 2 California 39536653 11439 29  
## 3 Florida 20984400 11059 53  
## 4 Texas 28304596 9756 34  
## 5 New York 19849399 7456 38  
## 6 Ohio 11658609 7275 62  
## 7 Georgia 10429379 6469 62  
## 8 Pennsylvania 12805537 6146 48  
## 9 North Carolina 10273419 6098 59  
## 10 Louisiana 4684333 5749 123  
## # ... with 41 more rows

# 주별 편차 고려하여 scale 조정  
quantiles1 <- quantile(b2$acc\_pop,  
 probs=seq(0, 1,length.out = 6+1))  
labels<- c()   
for(idx in 1:length(quantiles1)){  
 labels <- c(labels, paste0(round(quantiles1[idx], 2),  
 "-",  
 round(quantiles1[idx+1], 2)))  
}  
labels <- labels[1:length(labels)-1]   
# 사고 수 6개 범주로 나눔: acc\_pop\_quantiles  
b2$acc\_pop\_quantiles <- cut(b2$acc\_pop,  
 breaks = quantiles1,  
 labels = labels,  
 include.lowest = T)  
  
# Alaska 제외하고 시각화하였다.  
b %>% arrange(desc(acc\_pop)) %>% select(state, pop, acc, acc\_pop)

## # A tibble: 51 x 4  
## state pop acc acc\_pop  
## <chr> <int> <int> <dbl>  
## 1 District of Columbia 693972 2177 314  
## 2 Alaska 739795 1142 154  
## 3 Louisiana 4684333 5749 123  
## 4 Delaware 961939 1083 113  
## 5 Illinois 12802023 13600 106  
## 6 South Carolina 5024369 4906 98  
## 7 Mississippi 2984100 2542 85  
## 8 Tennessee 6715984 5650 84  
## 9 Missouri 6113532 4984 82  
## 10 Alabama 4874747 3804 78  
## # ... with 41 more rows

( text2<- b %>%   
 filter(acc\_pop >100, state != "Alaska") %>%   
 select(state, lat, long) )

## # A tibble: 4 x 3  
## state lat long  
## <chr> <dbl> <dbl>  
## 1 Delaware 39.3 -75.5  
## 2 District of Columbia 38.9 -77.0  
## 3 Illinois 40.3 -89.0  
## 4 Louisiana 31.2 -91.9

b2 %>%  
 ggplot(aes(long, lat)) +  
 geom\_polygon(aes(fill=acc\_pop\_quantiles, group=group),  
 colour="grey50", size=0.1) +  
 geom\_text(aes(long, lat, label=state), data=text2,  
 colour="grey70", fontface = "bold") +  
 coord\_equal() +  
 theme\_void() +  
 labs( title = "The number of accidents by states",  
 subtitle = "considering population size" ) +  
 scale\_fill\_viridis( option = "magma",  
 name = "The number\nof accident",  
 discrete = TRUE,  
 direction = -1 ) +  
 theme( plot.title = element\_text(hjust=0.5, colour="#4e4d47") ,  
 plot.subtitle = element\_text(hjust = 0.5, color = "#4e4d47") )

