# Midterm Project

*achyutganti*

*11/2/2017*

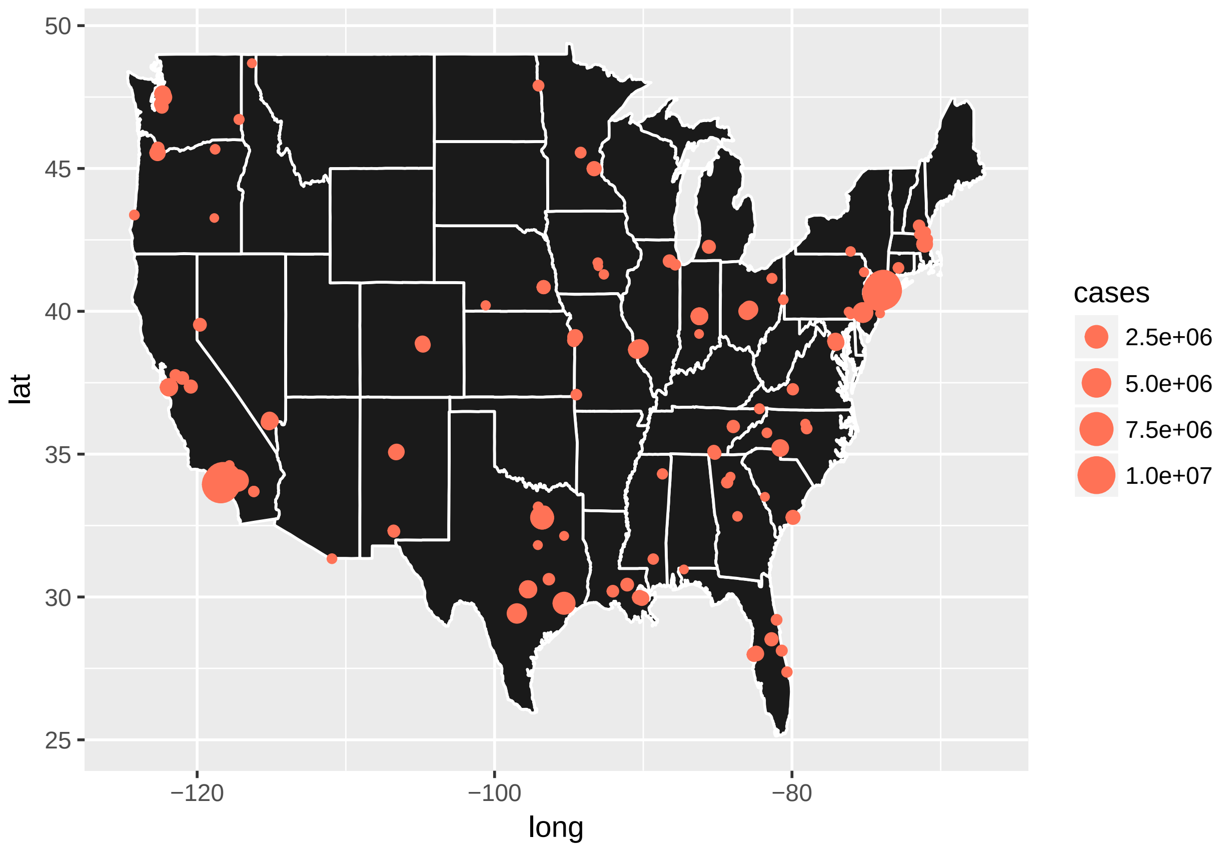
### Data Visualization in R

```
finaldata<- read.csv(’finaldata.csv’)
```

## 1Q Incorporating the population sizes and the three poverty variables on the US map.

### a - Population size

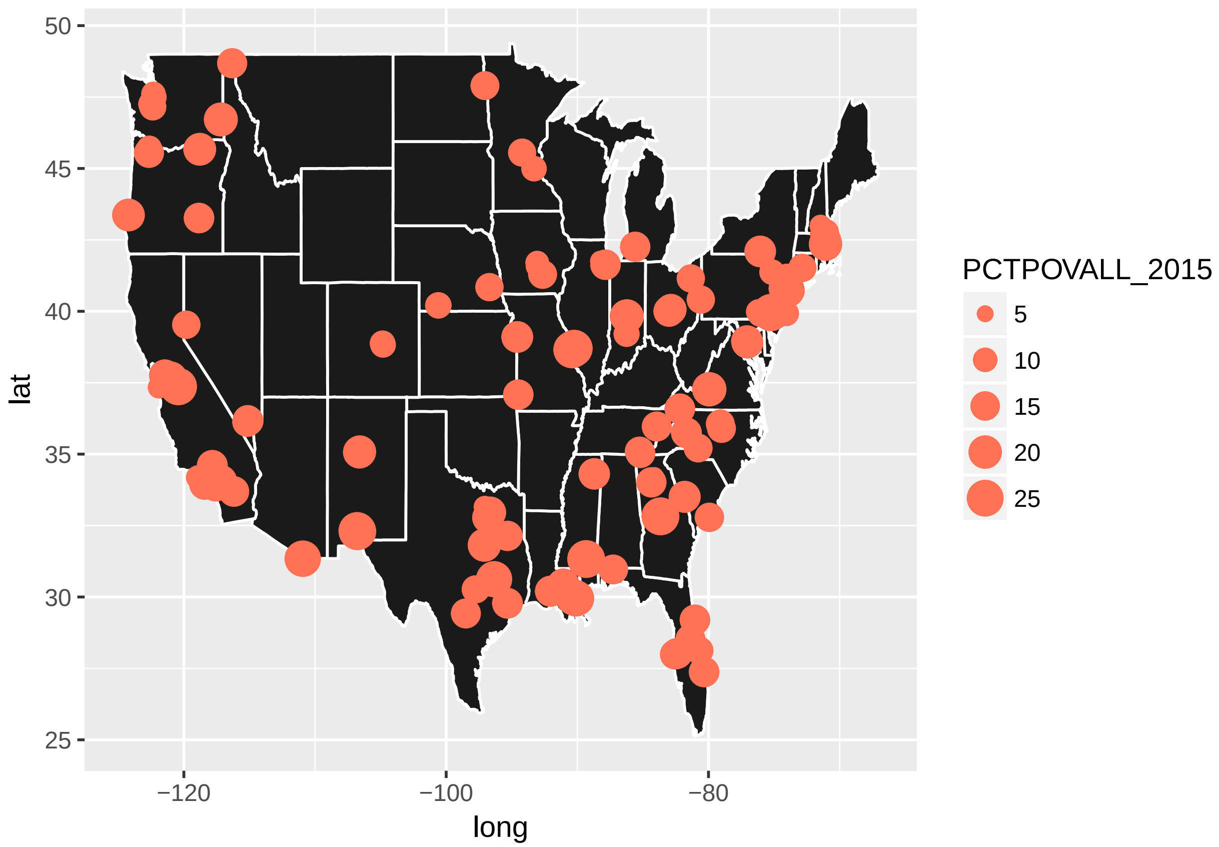
```
all_states<- map_data(’state’)
p<- ggplot()
p<- p+geom_polygon(data=all_states,aes(x=long,y=lat,group=group),color=’white’,fill=’grey10’)


#Mapping US population onto the map with respect to the sizes.
p<- p+geom_point(data=finaldata,aes(x=longitude,y=latitude,size=cases),color=’coral1’)
p
```



From the above plot we can conclude that two cities New york and California have the highest population sizes compared to the rest.

### b - Poverty variables (PCTPOVALL_2015)

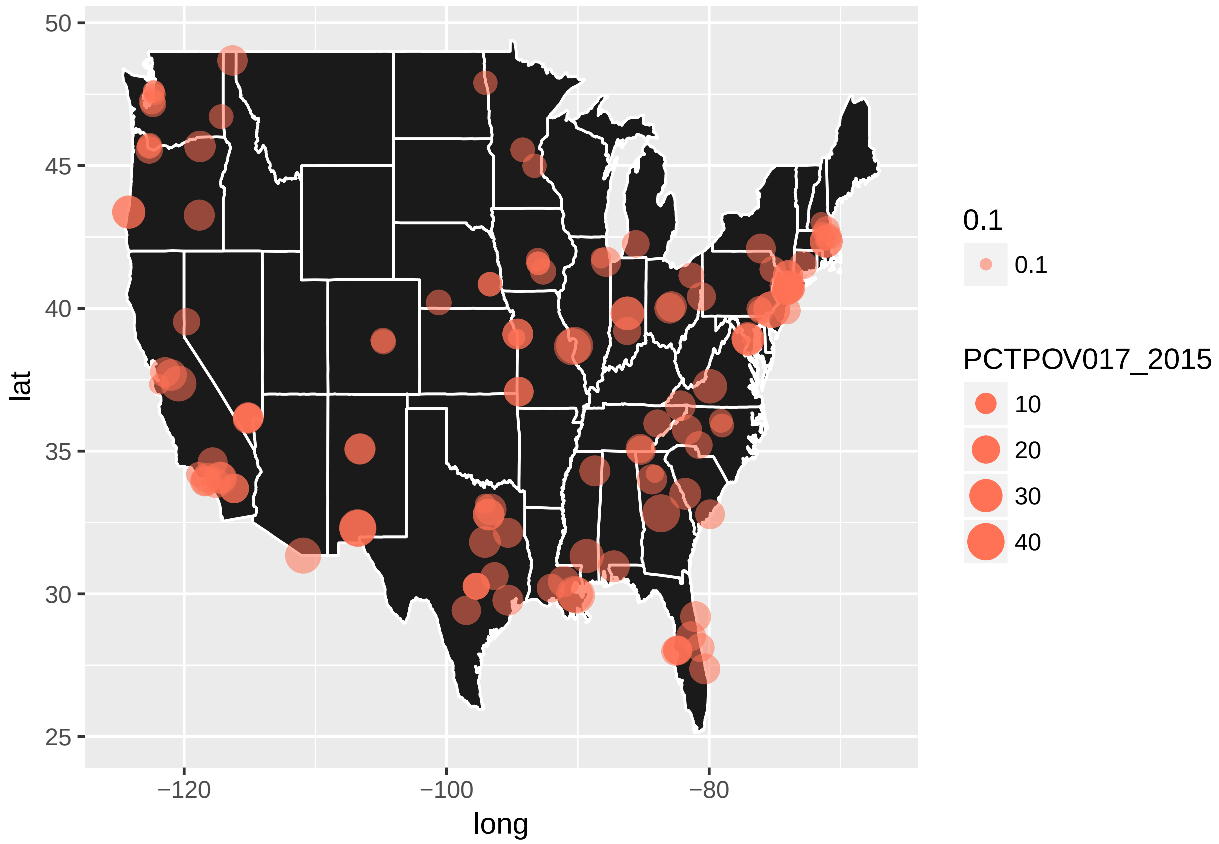
```
all_states<- map_data('state')
p<- ggplot()
p<- p+geom_polygon(data=all_states,aes(x=long,y=lat,group=group),color='white',fill='grey10')

p<- p+geom_point(data=finaldata,aes(x=longitude,y=latitude,size=PCTPOVALL_2015),color='coral1')
p
```



I don't see any clear pattern in the above US map. So, from our data and our plot the PCTOVALL_2015 value for all the counnties/ cities are approximately in the same magnitude.

### c - Poverty variables (PCTPOV017_2015)

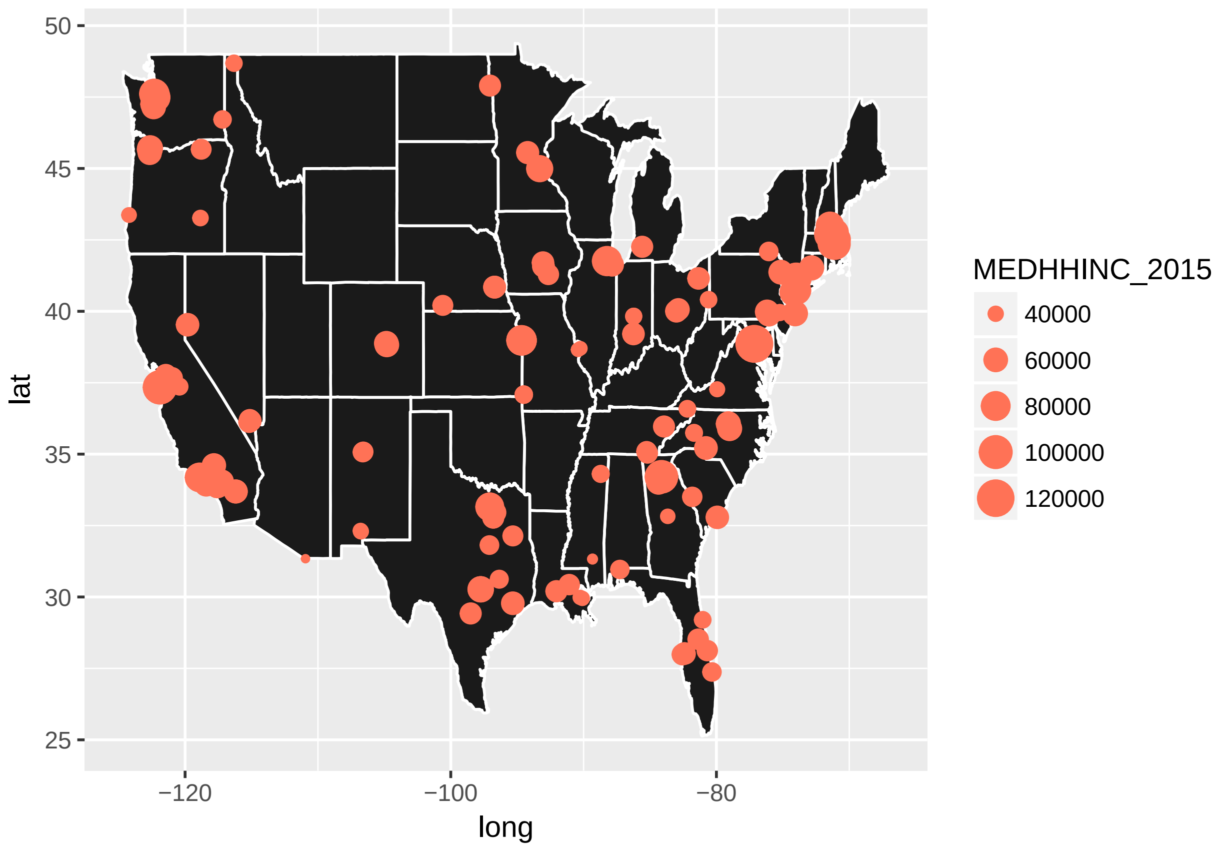
```
all_states<- map_data('state')
p<- ggplot()
p<- p+geom_polygon(data=all_states,aes(x=long,y=lat,group=group),color='white',fill='grey10')


p<- p+geom_point(data=finaldata,aes(x=longitude,y=latitude,size=PCTPOV017_2015,alpha=0.1),color='coral1
p
```



Looks like New York has the most poverty when compared to all the states. This plot doesn't reveal much except for that information. Because, the the poverty values were plotted with respect to their magnitude and they all look almost alike.

### d - Poverty variables ()

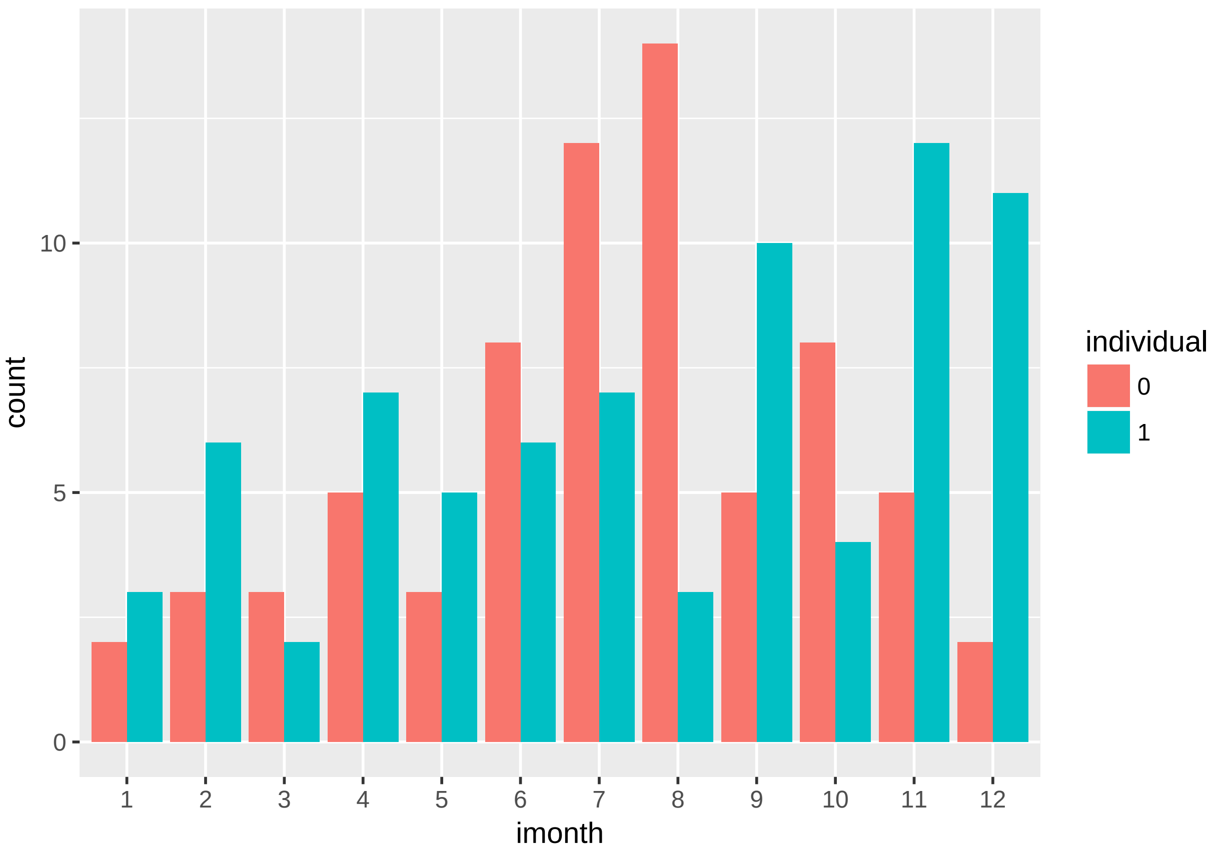
```
all_states<- map_data('state')
p<- ggplot()
p<- p+geom_polygon(data=all_states,aes(x=long,y=lat,group=group),color='white',fill='grey10')


p<- p+geom_point(data=finaldata,aes(x=longitude,y=latitude,size=MEDHHINC_2015,color=provstate),color='c
p
```



The median house hold income of New York, California and some parts of Washington seem to be higher than the rest of the states. Florida has very low median income. Also the middle parts of the US has average income. They are no where near the New york and California incomes.

## 2Q Making a plot of incidents by month by variable individual.

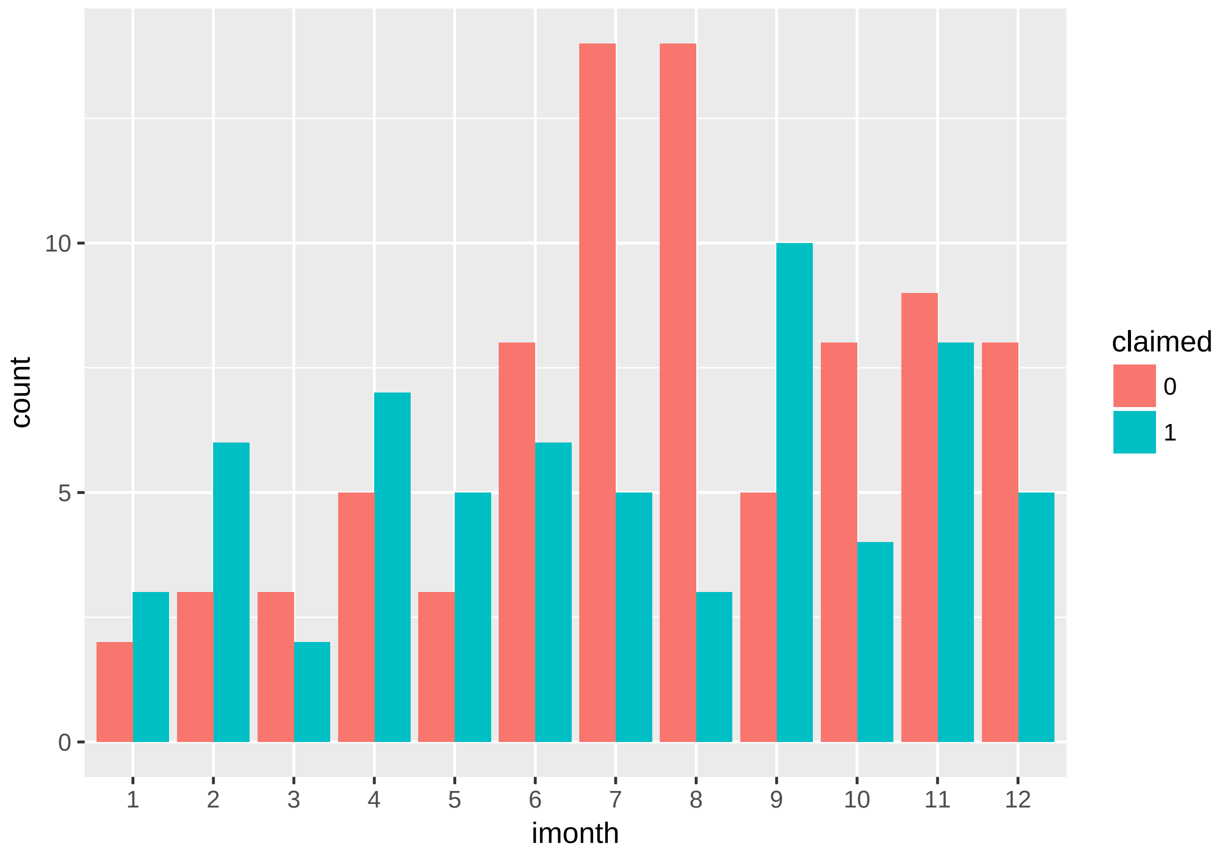
```
finaldata$individual<- as.factor(finaldata$individual)
finaldata$imonth<- as.factor(finaldata$imonth)
finaldata%>%group_by(imonth)%>%
  ggplot() + geom_bar(aes(x=imonth,fill = individual), position = "dodge")
```



The above plot explains if the attacks were identified as being carried out by the individual(*1*) or a group (*0*) The attacks carried out by individual are more than that of attacks carried out by groups. In 7 out of 12 months, the attacks carried out by individuals are more. Also, in July and August, the perpetrators were affiliated with a group or organization.

### Claimed -vs- imonth.

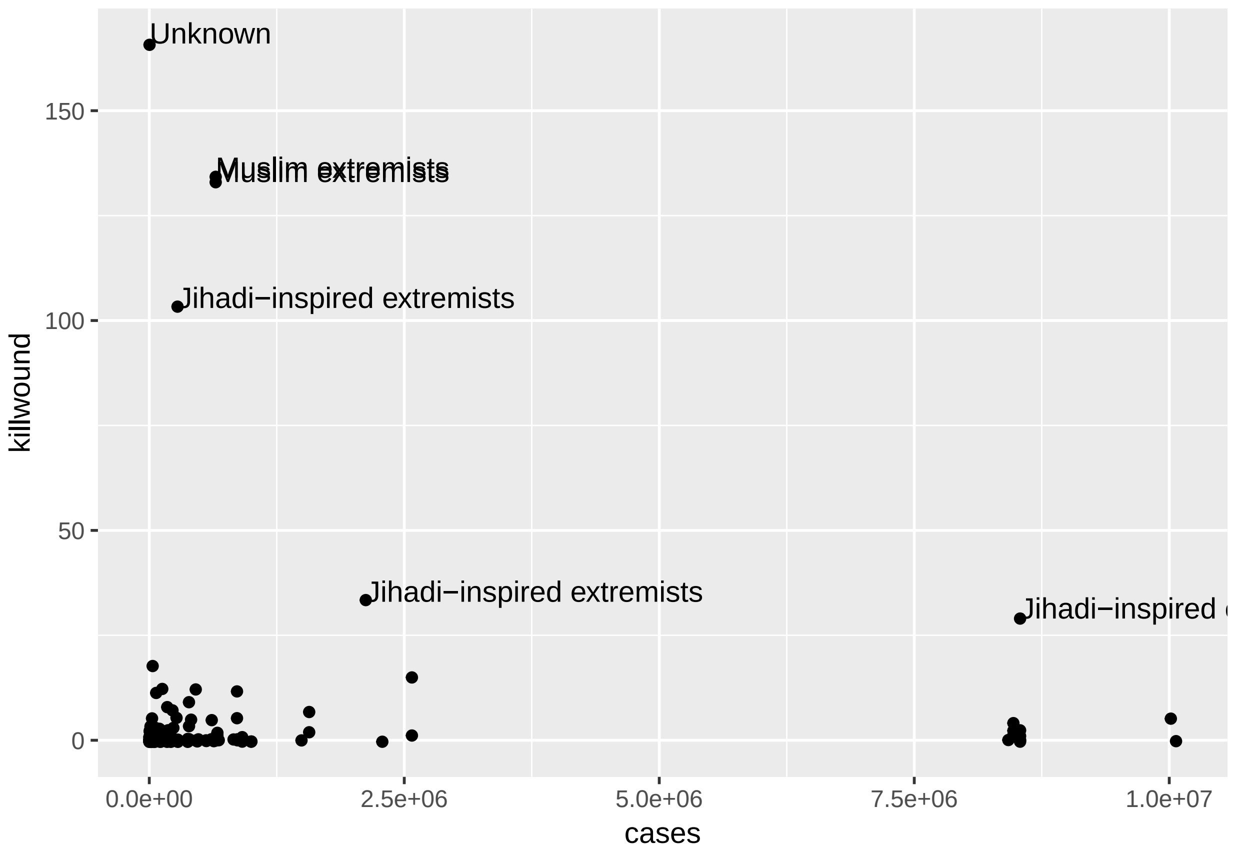
```
finaldata$claimed<- as.factor(finaldata$claimed)
finaldata%>%group_by(imonth)%>%
  ggplot() + geom_bar(aes(x=imonth,fill = claimed), position = "dodge")
```



This plot shows whether a group of attackers claimed responsibility for the attack or not. Where *1* is claimed and *0* is not. Surprisingly, in July and August, the perpetrators were identified as a group. But they did not claim responsible for the attack.

## 3Q Relationship between killed and wounded combined with populations(cases)

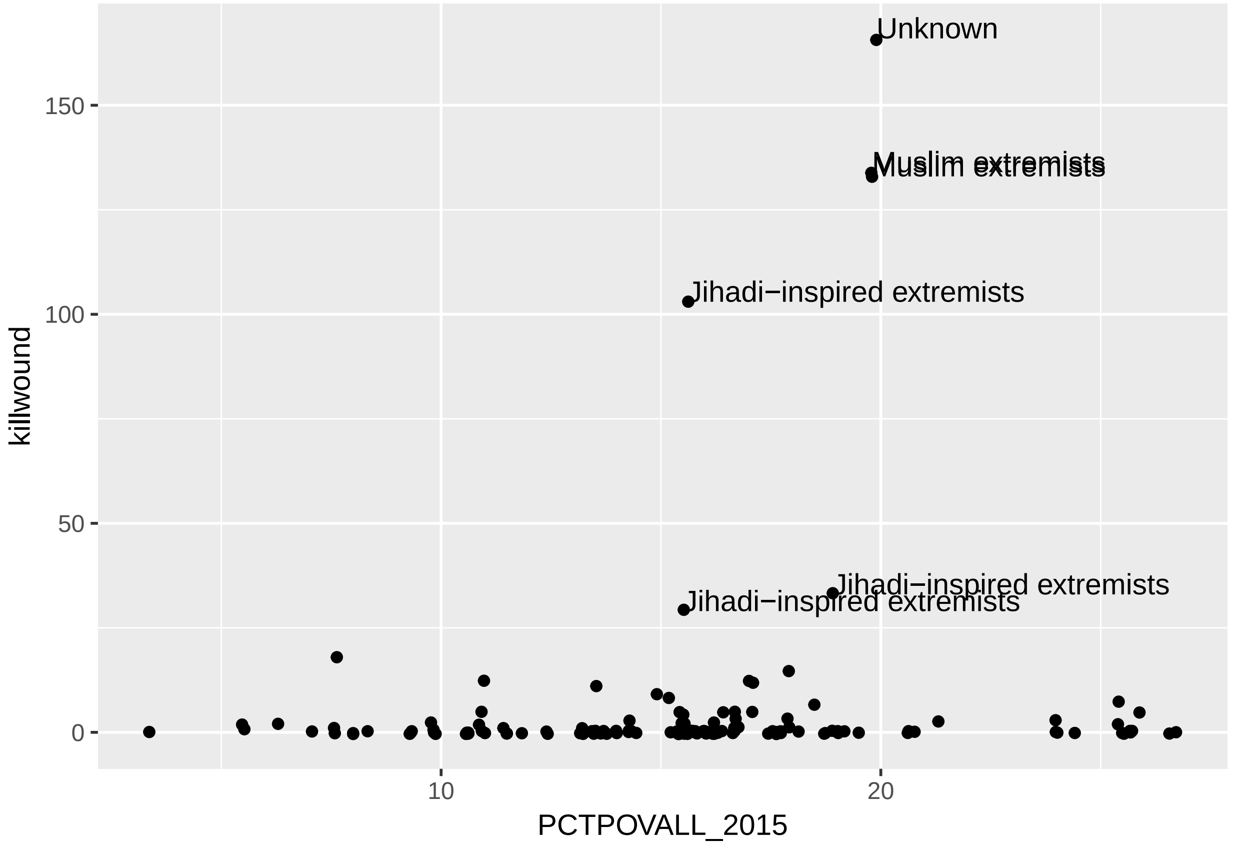
```
combined <- finaldata%>%mutate(killwound = nkill+nwound)

ggplot(combined,aes(x= cases, y=killwound,label=gname))+geom_point(position='jitter')+
  geom_text(aes(label=ifelse(killwound>25,as.character(gname),'')),hjust=0,vjust=0)
```



From the above plot, I can say that Muslim Extremists and Jihadi-inspired extremists were involved in most kills and wounds of poeple. And we can also conclude that these groups attacked places where the population was less.

### Poverty variable PCTOVALL_2015 with nkill and nwound combined.

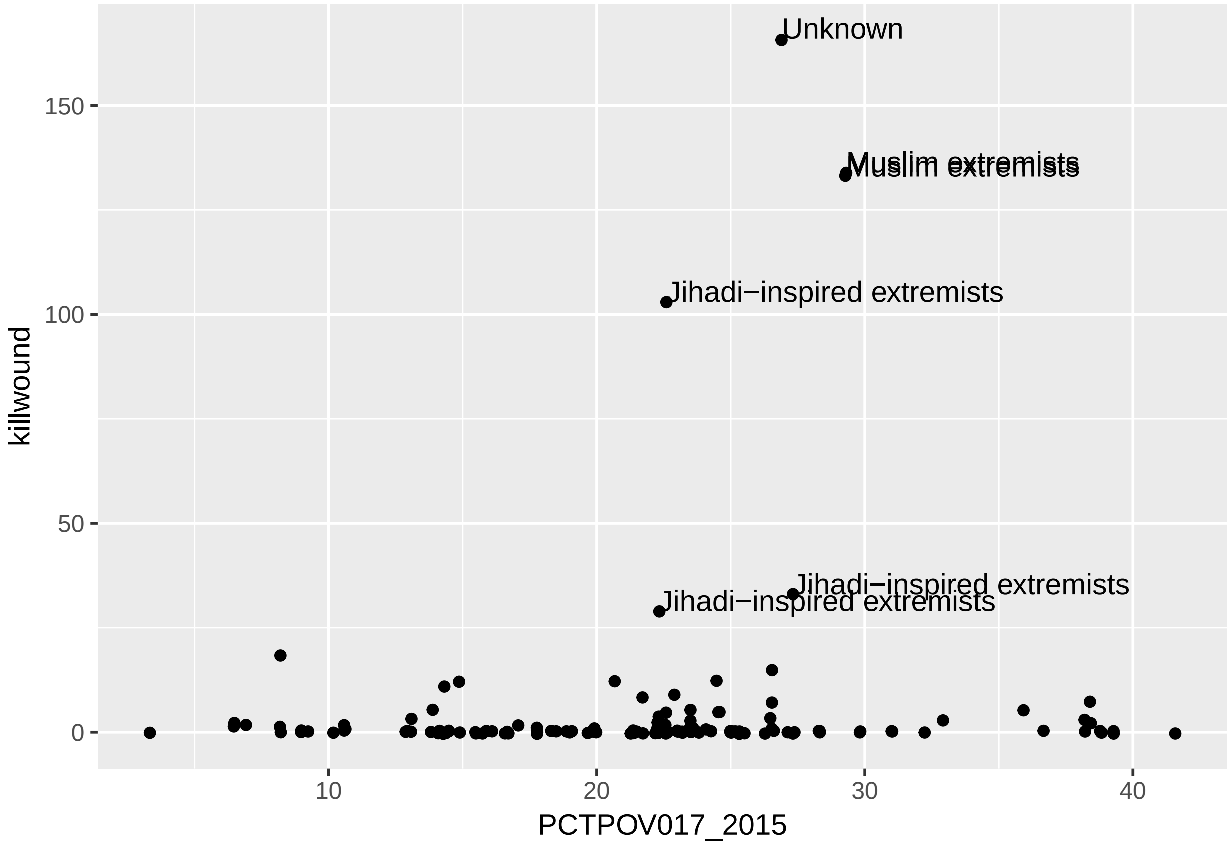
```
ggplot(combined,aes(x= PCTPOVALL_2015, y=killwound,label=gname))+geom_point(position='jitter')+
  geom_text(aes(label=ifelse(killwound>25,as.character(gname),'')),hjust=0,vjust=0)
```



In between the PCTOVALL_2015 value of 15 and 20, we can see most kills nd wounds. Even here Muslim Extremisits and Jihadi inspired people injured most people. Surprisingly, they injured people where poverty levels are a bit high. So, we ca coclude that less denser areas Poorer people tend to get attacked more often than the rich.

### Poverty variable PCTPOV017_2015 with nkill and combined.

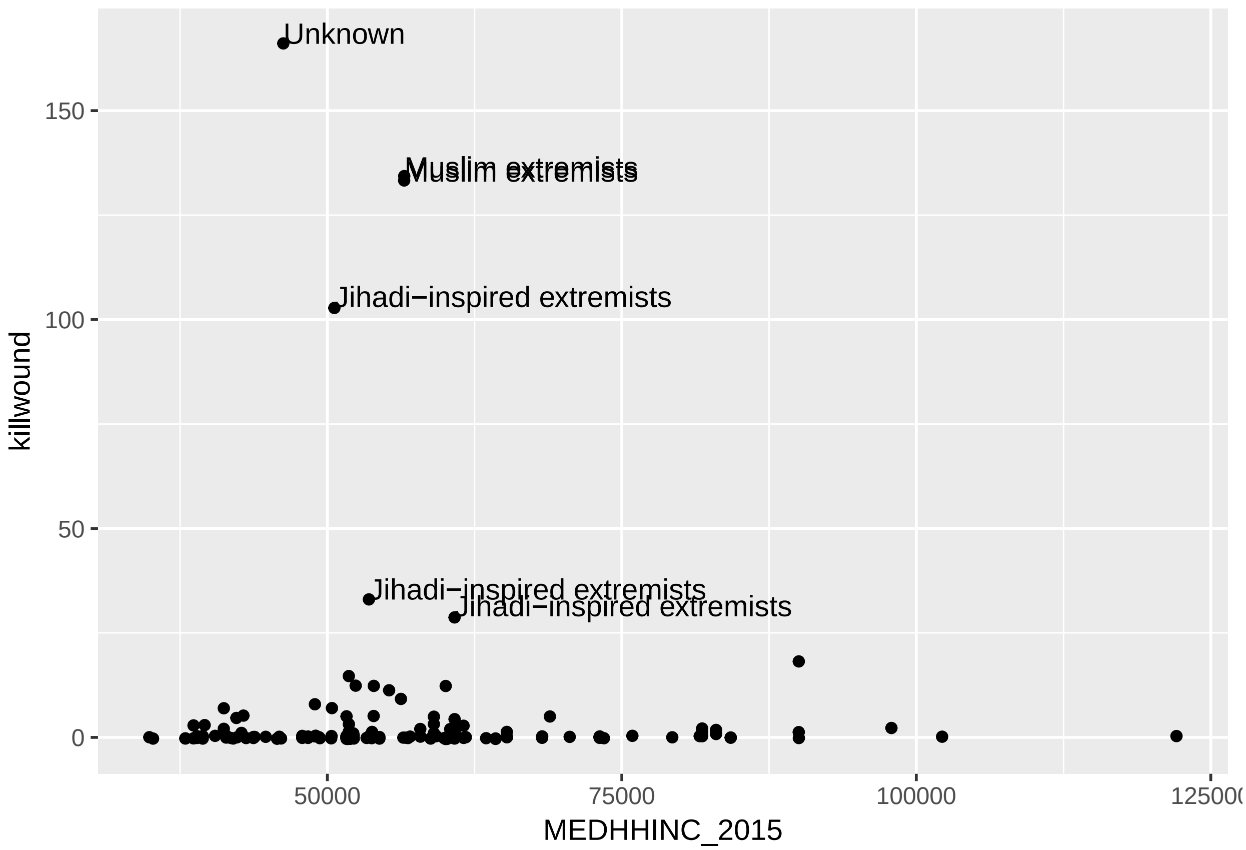
```
ggplot(combined,aes(x= PCTPOV017_2015, y=killwound,label=gname))+geom_point(position='jitter')+
  geom_text(aes(label=ifelse(killwound>25,as.character(gname),'')),hjust=0,vjust=0)
```



Almost all the datapoints are in between 0 and 25 except for some. All those large number of attacks fall under the category of PCTPOV017_2015 value in between 20 and 30. Also, by looking at the plot we can conclude that Muslim and Jihadi extremists were involved in most kills and woundings.

### Poverty variable Income with nkill and wounded

```
ggplot(combined,aes(x= MEDHHINC_2015, y=killwound,label=gname))+geom_point(position='jitter')+
  geom_text(aes(label=ifelse(killwound>25,as.character(gname),'')),hjust=0,vjust=0)
```



As median income increases there are less kills and woundings. Most of the kills happened below 60000$. Surprisingly, even here Muslim and Jihadi Extremists killed most people.

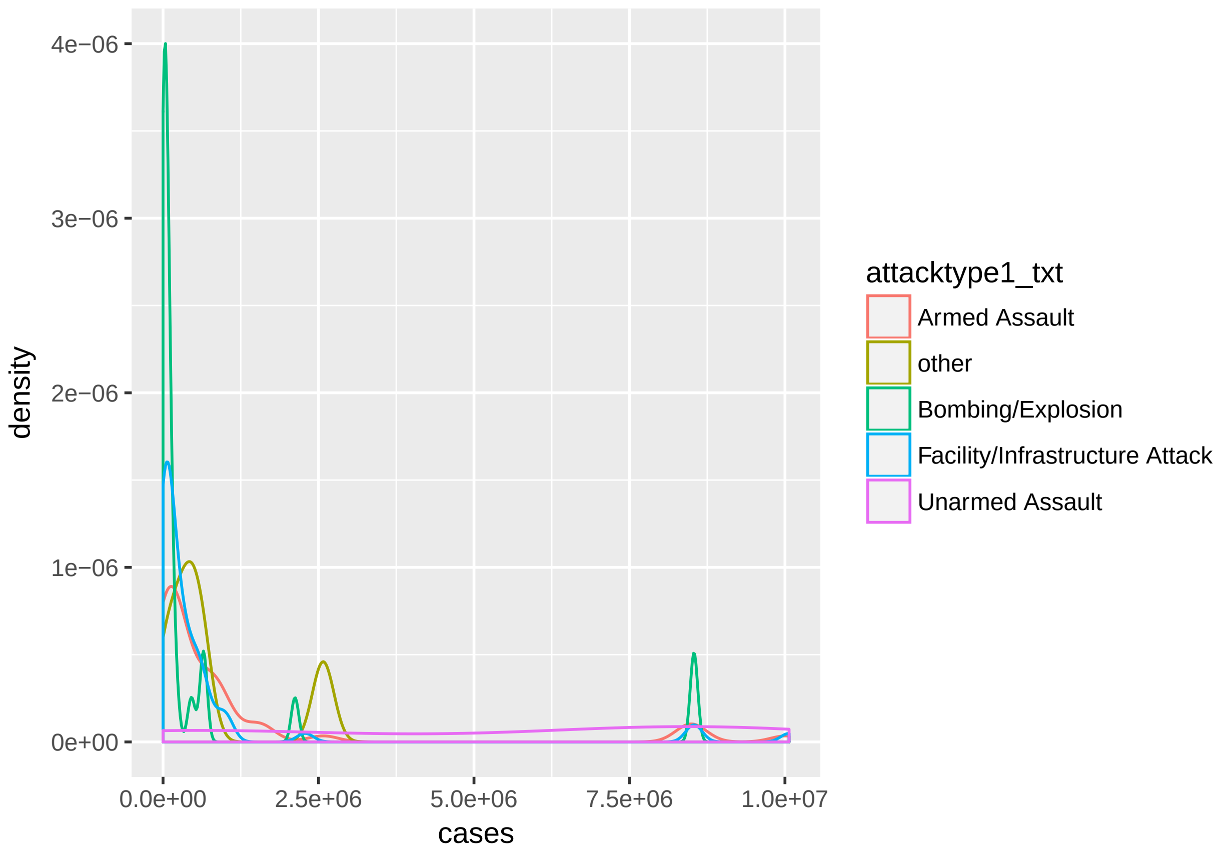
## 4Q Attacktype with population (cases) and teh three other poverty variables.

```
finaldata$attacktype1_txt <- as.factor(finaldata$attacktype1_txt)
attack_types<-finaldata%>%group_by(attacktype1_txt)%>%summarise(count=n())

collapsed<- finaldata%>%mutate(attacktype1_txt=fct_collapse(attacktype1_txt,
                                                  other=c('Assassination','Hijacking','Hostage Taking (Ba
```

I have summarised the attacktype1__txt variable after grouping it. I have observed that three levels had values less than 3. So i have collapsed them into a single Other category.

### Attacktype with population (cases)

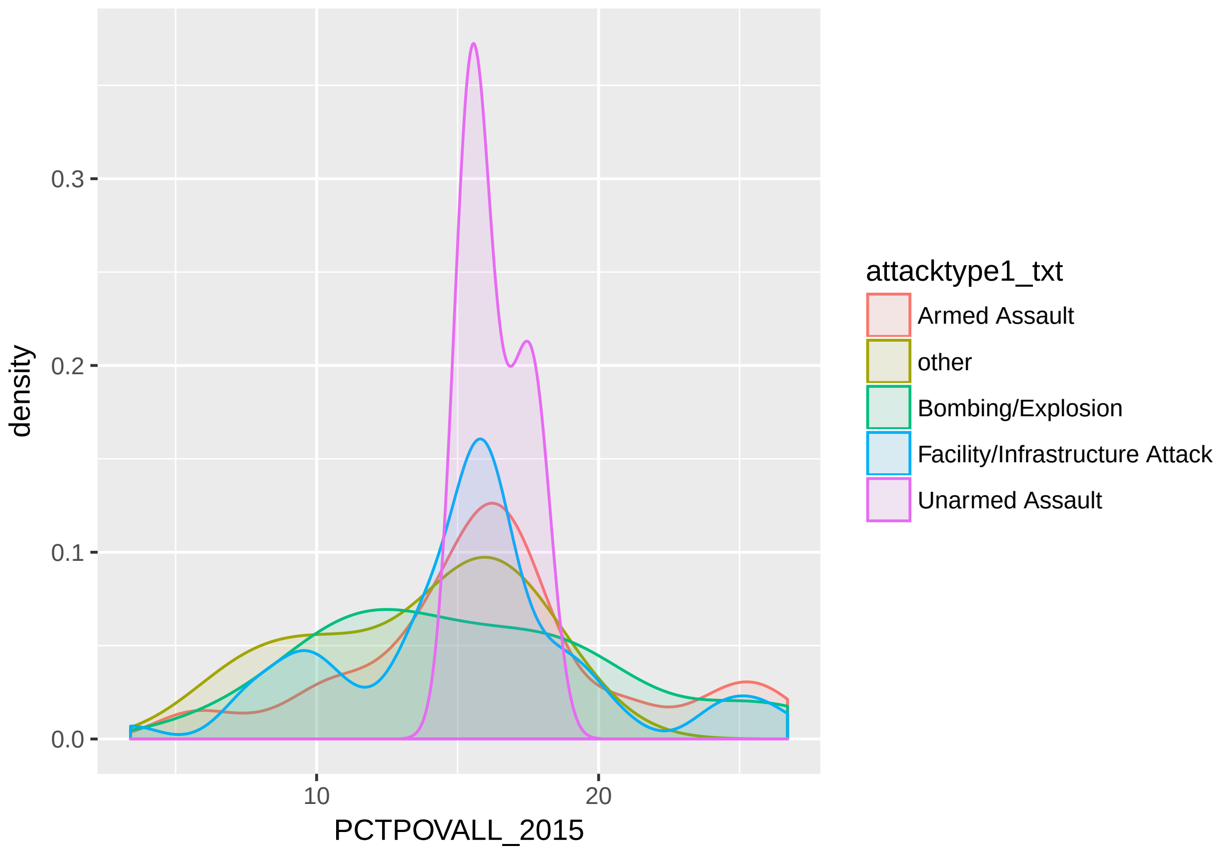
```
ggplot(collapsed, aes(x = cases, color = attacktype1_txt)) + geom_density()
```



In the above density plot, we can observe that most of the heavier attacks happened in areas where there was lesser population with bombing explosion killing most people as expected.

## Relationship between Poverty variable PCTPOVALL_2015 and attacktype1_txt

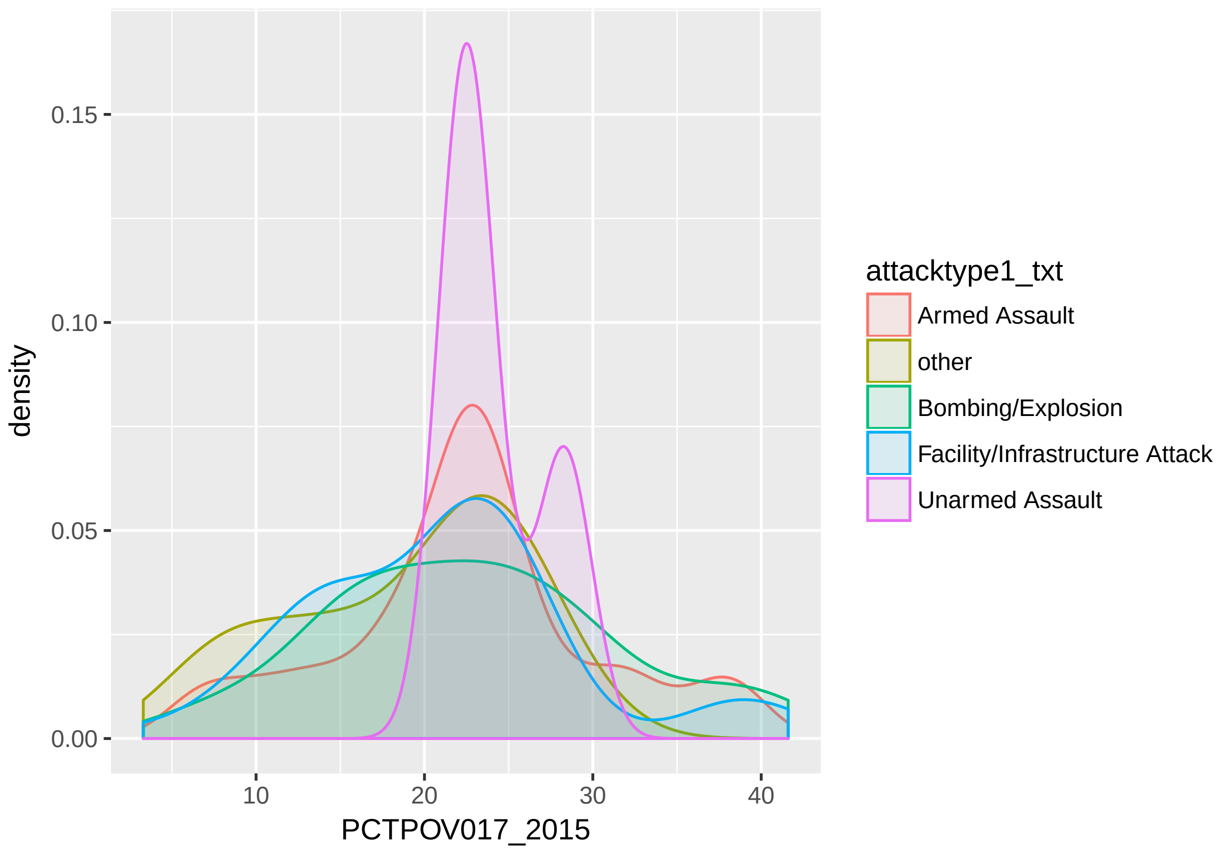
```
ggplot(collapsed, aes(x =PCTPOVALL_2015 , color = attacktype1_txt,fill=attacktype1_txt)) + geom_density
```



*Conclusion*-The distribution of Unarmed Assaults and Faculty/Infrastructure Attacks is more than other attacks. So these kind of attacks took most lives when compared to others.

### Relationship between Poverty variable PCTPOV017_2015 and attacktype1_txt

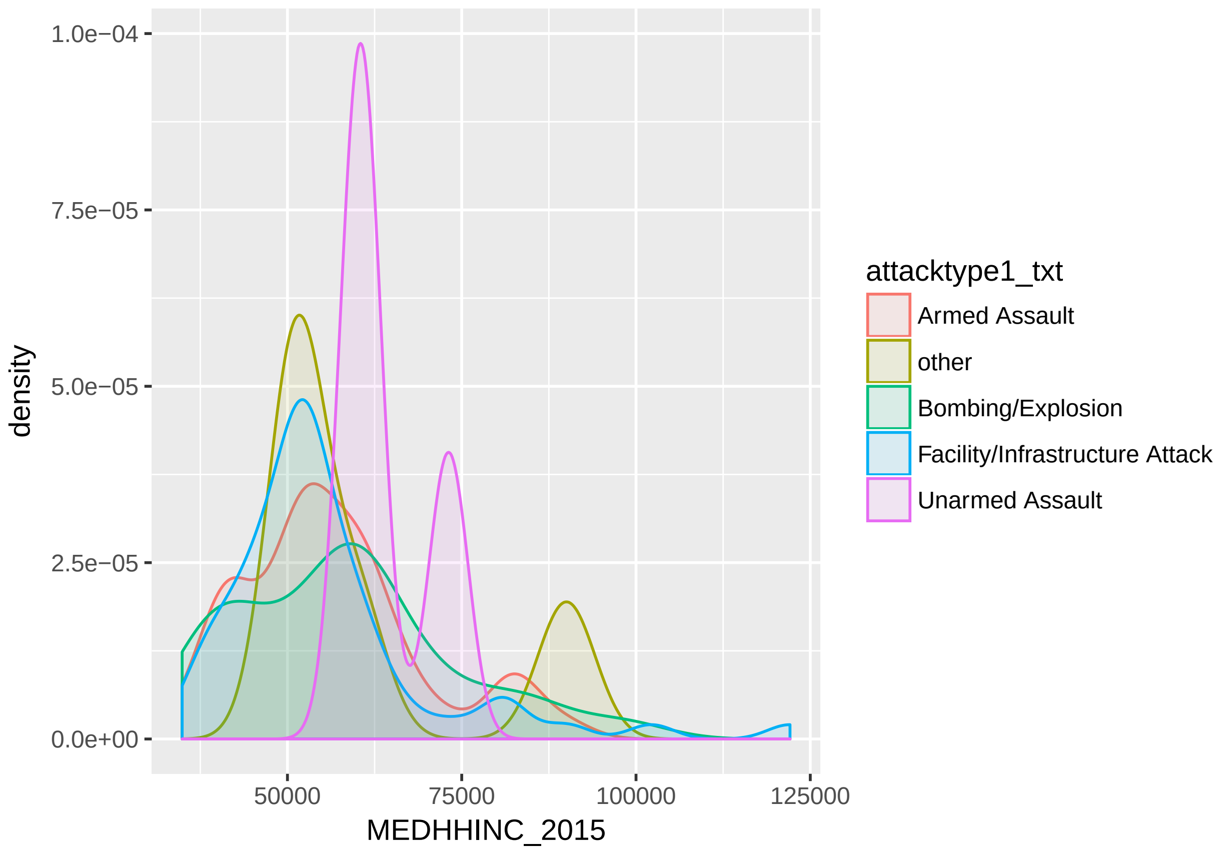
```
ggplot(collapsed, aes(x =PCTPOV017_2015 , color = attacktype1_txt,fill=attacktype1_txt)) + geom_density
```



*Conclusion* - Unarmed attacks has a taller peak but the Bombing explosions and Infrastructures attacks have fatter ends. They may be lethal when compared to other ones.

### Relationship between Poverty variable MEDHHINC_2015 and attacktype1_txt

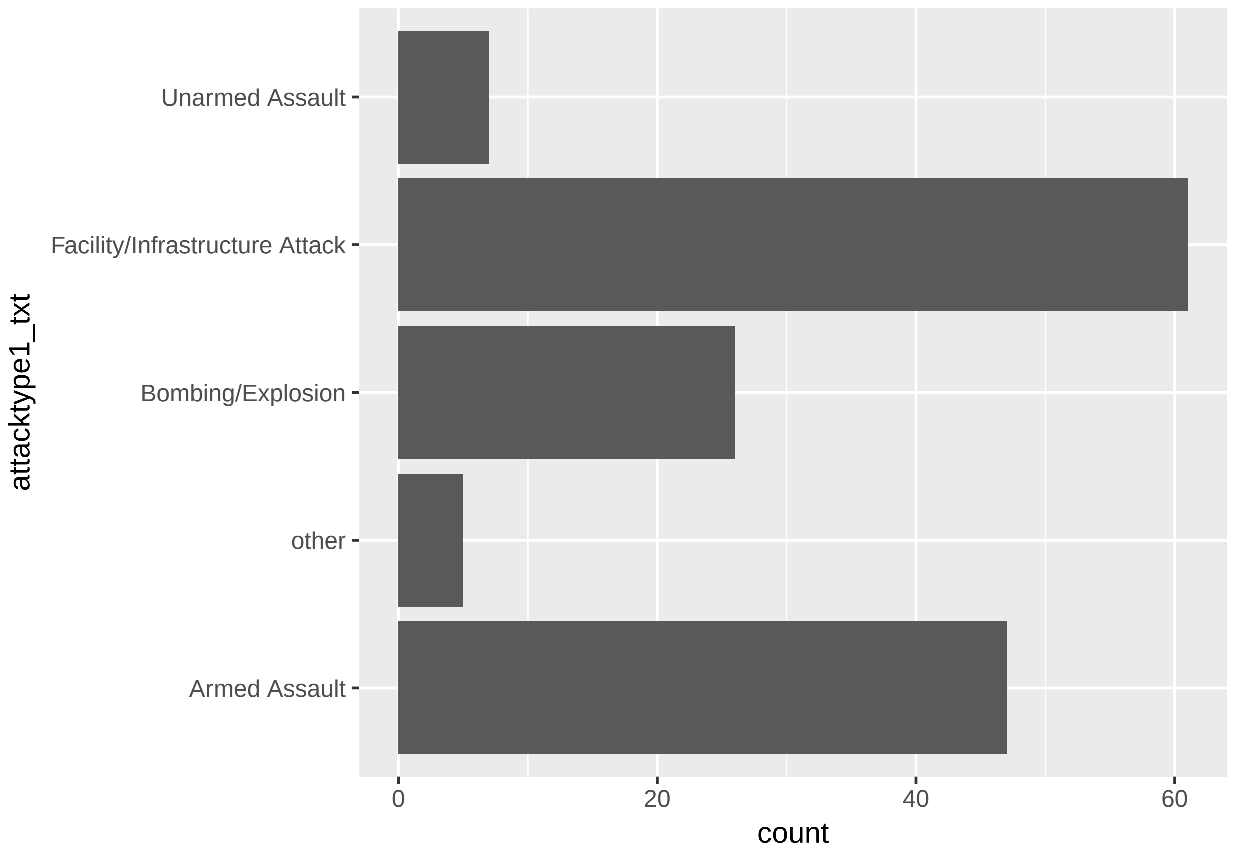
```
ggplot(collapsed, aes(x =MEDHHINC_2015 , color = attacktype1_txt,fill=attacktype1_txt)) + geom_density(a
```



As discussed earlier, lesser median income poeple were attacked more than the higher median income people. As we can see, all the attacks killed or wounded most people of income lesser than 60000$

## Exploring the relationship between attacktype and property.

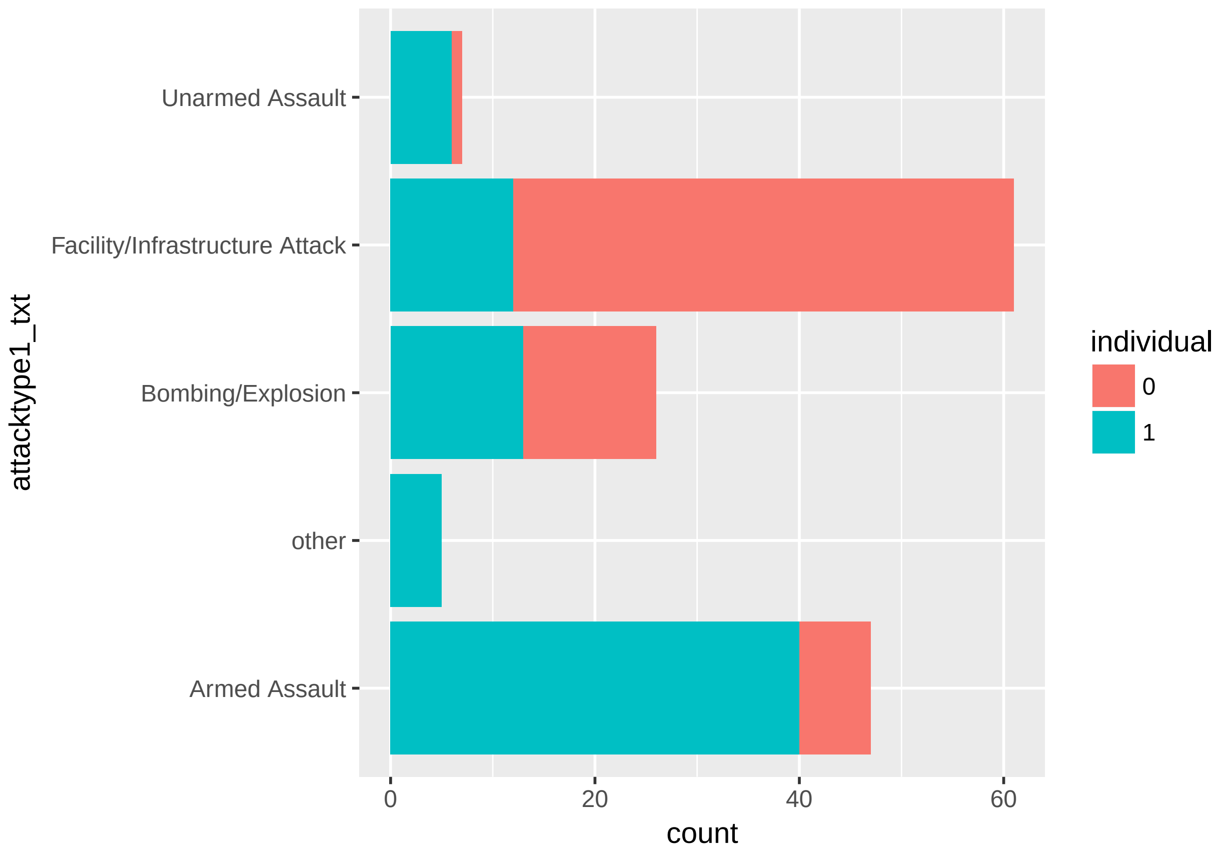
```
finaldata$property<-as.factor(finaldata$property)
ggplot(collapsed, aes(x = attacktype1_txt, fill = property)) + geom_bar()+coord_flip()
```



The above plot shows which attack led to more property damage and which did not. If property is 1, then we can say there is a property damage. If it is 0, then there is not. Intuitively, Facility/ Infrastuctural Attacks led to more property damage, while armed attacks led to less property damage.

Relationship between attacktype and individual

```
ggplot(collapsed, aes(x = attacktype1_txt, fill = individual)) + geom_bar()+coord_flip()
```



Here the plot shows which attacks were carried out mostly by individuals and which ones by groups. 1 indicates that the attacks were done by groups. whereas 0 indicates a group attack. Armed assaults were mostly done by individuals whereas Infrastructure attacks were done by groups.