Download the Washington Post database of US police shootings. Which of the following techniques might be useful in addressing questions arising from this data set?

- classification
- regression
- cluster analysis

Link for the Washington Post Database of US Police Shootings - View Article Download Dataset

Washington Post released a database in which they recorded every fatal shooting in United States by police officer in line of duty since January 1 2015.

The Dataset `fatal-police-shootings-data.csv` is in the csv format and contains following `variables`.

1. `id` - unique identifier of each victim.
2. `name` - name of the person who got shot.
3. `date` - Date of shooting in YYYY-MM-DD format.
4. `manner_of_death` - This field tells us the way person was killed and only has two values (a) Shot and (b) Shot and Tasered.
5. `Armed` - It tells us the nature of victim whether they were armed, unarmed or unknown. According to washington post official Github account

   - `undetermined`: it is not known whether or not the victim had a weapon
   - `unknown`: the victim was armed, but it is not known what the object was
   - `unarmed`: the victim was not armed

6. `Age` - defines the age of a person shot.
7. `Gender` - Gender of a person Male or Female or Unknown in some cases.
8. `Race` - Race defines the race of person shot where
   - W: White, non-Hispanic
   - B: Black, non-Hispanic
   - A: Asian
   - N: Native American
   - H: Hispanic
   - O: Other
   - None: unknown
9. `City` - The municipality where shooting happened and in some cases nearest county is listed.
10. `State` - Name of the satate where it happened.
11. `Signs of Mental illness` - Whether the victim had Mental Illness during the time period when shooting happened. True or False.

12. `threat_level` : According to washington post official Github account The threat_level column was used to flag incidents for the story by Amy Brittain in October 2015. http://www.washingtonpost.com/sf/investigative/2015/10/24/on-duty-under-fire/ As described in the story, the general criteria for the attack label was that there was the most direct and immediate threat to life. That would include incidents where officers or others were shot at, threatened with a gun, attacked with other weapons or physical force, etc. The attack category is meant to flag the highest level of threat. The other and undetermined categories represent all remaining cases. Other includes many incidents where officers or others faced significant threats.
13. `Flee` - Whether the victim was trying to flee from officer using
    - Car.
    - Foot.
    - Not Fleeing.
14. `body_camera` - Indiactes that whether the officer was wearing a body cam or not.
15. `latitude and longitude` : According to washington post official Github account `latitude and longitude` is the location of the shooting expressed as WGS84 coordinates, geocoded from addresses. The coordinates are rounded to 3 decimal places, meaning they have a precision of about 80-100 meters within the contiguous U.S.

16. `is_geocoding_exact` : According to washington post official Github account `is_geocoding_exact` reflects the accuracy of the coordinates. true means that the coordinates are for the location of the shooting (within approximately 100 meters), while false means that coordinates are for the centroid of a larger region, such as the city or county where the shooting happened.

Answer

- The question demands a conclusion on which technique (Regression, Classification, Clustering) would be helpful in addressing the question arising from this dataset.

- Although, before concluding that, Let's perform Exploratory Data Analysis on the Data set to have a better understanding of the dataset and also it will eventually help to figure out which technique will suit the best.

Using R in python environment.

We can simply visit this link with a language parameter and use directly Link

Library and R setting

In [ ]:

```
library(tidyverse)  # A widdely used R Library for Visualization and Manipulation
options(repr.plot.width = 8, repr.plot.height = 4, repr.plot.res = 200) # Setting R environment
```

Loading Dataset

In [ ]:

```
data <- read.csv('/content/fatal-police-shootings-data.csv')   #Load the file and save it to th
```

Understand the Structure of the data

It will help us observe the different predictors available to us

In [ ]:

```
str(data)
```

```
'data.frame':	7729 obs. of  17 variables:
 $ id                     : int  3 4 5 8 9 11 13 15 16 17 ...
 $ name                   : chr  "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew 
Hoffman" ...
 $ date                   : chr  "2015-01-02" "2015-01-02" "2015-01-03" "2015-01-04" ...
 $ manner_of_death        : chr  "shot" "shot" "shot and Tasered" "shot" ...
 $ armed                  : chr  "gun" "gun" "unarmed" "toy weapon" ...
 $ age                    : int  53 47 23 32 39 18 22 35 34 47 ...
 $ gender                 : chr  "M" "M" "M" "M" ...
 $ race                   : chr  "A" "W" "H" "W" ...
 $ city                   : chr  "Shelton" "Aloha" "Wichita" "San Francisco" ...
 $ state                  : chr  "WA" "OR" "KS" "CA" ...
 $ signs_of_mental_illness: chr  "True" "False" "False" "True" ...
 $ threat_level           : chr  "attack" "attack" "other" "attack" ...
 $ flee                   : chr  "Not fleeing" "Not fleeing" "Not fleeing" "Not fleeing" ...
 $ body_camera            : chr  "False" "False" "False" "False" ...
 $ longitude              : chr  "-123.122" "-122.892" "-97.281" "-122.422" ...
 $ latitude               : chr  "47.247" "45.487" "37.695" "37.763" ...
 $ is_geocoding_exact     : chr  "True" "True" "True" "True" ...
```

> If we observe the above output then we can notice that we have 13 CHARACTER value, 2 INTEGER values (`id` and `age`) and 2 NUMERICAL values (`longitude` and `latitude`)
>
> Looking at the structure of the data, we can expect that we won't be using Regression analysis as the integer value consist only `id` and `age` and rest `longitude` and `latitude` which is mostly used for confirming the geo location when shooting happened. Therefore, we can state that We won't be using the Regression in this Dataset.

Top 5 Values of Data

In [ ]:

```
head(data,5)
```

A data.frame: 5 × 17

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_ment
	<int>	<chr>	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	
1	3	Tim Elliot	2015-01-02	shot	gun	53	M	A	Shelton	WA	
2	4	Lewis Lee Lembke	2015-01-02	shot	gun	47	M	W	Aloha	OR	
3	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23	M	H	Wichita	KS	
4	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32	M	W	San Francisco	CA	
5	9	Michael Rodriguez	2015-01-04	shot	nail gun	39	M	H	Evans	CO	

Check for missing values and Count all the missing values in entire dataframe.

In [ ]:

```
sapply(data, function(x) sum(is.na(x)))
```

```
id: 0 name: 0 date: 0 manner_of_death: 0 armed: 0 age: 482 gender: 0 race: 0 city: 0 state: 0
 signs_of_mental_illness: 0 threat_level: 0 flee: 0 body_camera: 0 longitude: 0 latitude: 0
 is_geocoding_exact: 0
```

We can Observe from the above values that the `age` has 482 missing values. We can try to remove these missing values and replace it with a mean age. Although age is crucial criteria in determining the number of people killed in a particular age group.

`Longitude and Latitude` has 837 missing values each, these values are used to determine the exact geolocation of the person where he/she was shot. Even if we are trying to know the state or municipality wise data of people killed, we can work our way around other available column such as `city` or `state`.

We won't be handling missing values of `longitude` and `latitude` as that would make no sense replacing them with any random numbers.

Replacing missing values in `age` column

In [ ]:

```
data$age[is.na(data$age)] = mean(data$age, na.rm=TRUE)

sapply(data, function(x) sum(is.na(x)))
```

```
id: 0 name: 0 date: 0 manner_of_death: 0 armed: 0 age: 0 gender: 0 race: 0 city: 0 state: 0
 signs_of_mental_illness: 0 threat_level: 0 flee: 0 body_camera: 0 longitude: 0 latitude: 0
 is_geocoding_exact: 0
```

In [ ]:

```
head(data)
```

A data.frame: 6 × 17

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_ment
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	
1	3	Tim Elliot	2015-01-02	shot	gun	53	M	A	Shelton	WA	
2	4	Lewis Lee Lembke	2015-01-02	shot	gun	47	M	W	Aloha	OR	
3	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23	M	H	Wichita	KS	
4	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32	M	W	San Francisco	CA	

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_ment
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	
5	9	Michael Rodriguez	2015-01-04	shot	nail gun	39	M	H	Evans	CO	
6	11	Kenneth Joe Brown	2015-01-04	shot	gun	18	M	W	Guthrie	OK	

Summary of the data

In [ ]:

```
summary(data)
```

```
      id            name               date           manner_of_death   
 Min.   :   3   Length:7729        Length:7729        Length:7729       
 1st Qu.:2167   Class :character   Class :character   Class :character  
 Median :4293   Mode  :character   Mode  :character   Mode  :character  
 Mean   :4268                                                           
 3rd Qu.:6358                                                           
 Max.   :8406                                                           
    armed                age            gender              race          
 Length:7729        Min.   : 2.00   Length:7729        Length:7729       
 Class :character   1st Qu.:28.00   Class :character   Class :character  
 Mode  :character   Median :36.00   Mode  :character   Mode  :character  
                    Mean   :37.18                                        
                    3rd Qu.:45.00                                        
                    Max.   :92.00                                        
     city              state           signs_of_mental_illness
 Length:7729        Length:7729        Length:7729            
 Class :character   Class :character   Class :character       
 Mode  :character   Mode  :character   Mode  :character       
                                                              
                                                              
                                                              
 threat_level           flee           body_camera          longitude        
 Length:7729        Length:7729        Length:7729        Length:7729       
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                            
                                                                            
                                                                            
   latitude         is_geocoding_exact
 Length:7729        Length:7729       
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

Let's Try to visualise our data as available in the ARTICLE and then we can discuss the technique more suitable for the Washington Post Police Shootings Database.

1. People killed based on their gender.
2. People killed based on their Race.

3. People who showed Signs of Mental illness with respect to their race.
4. Victims According to their Age group.
5. Victims Killed by State.
6. Threat Level on the basis of Race.
7. Police shootings based on Year since 2015.

1. People killed based on their gender

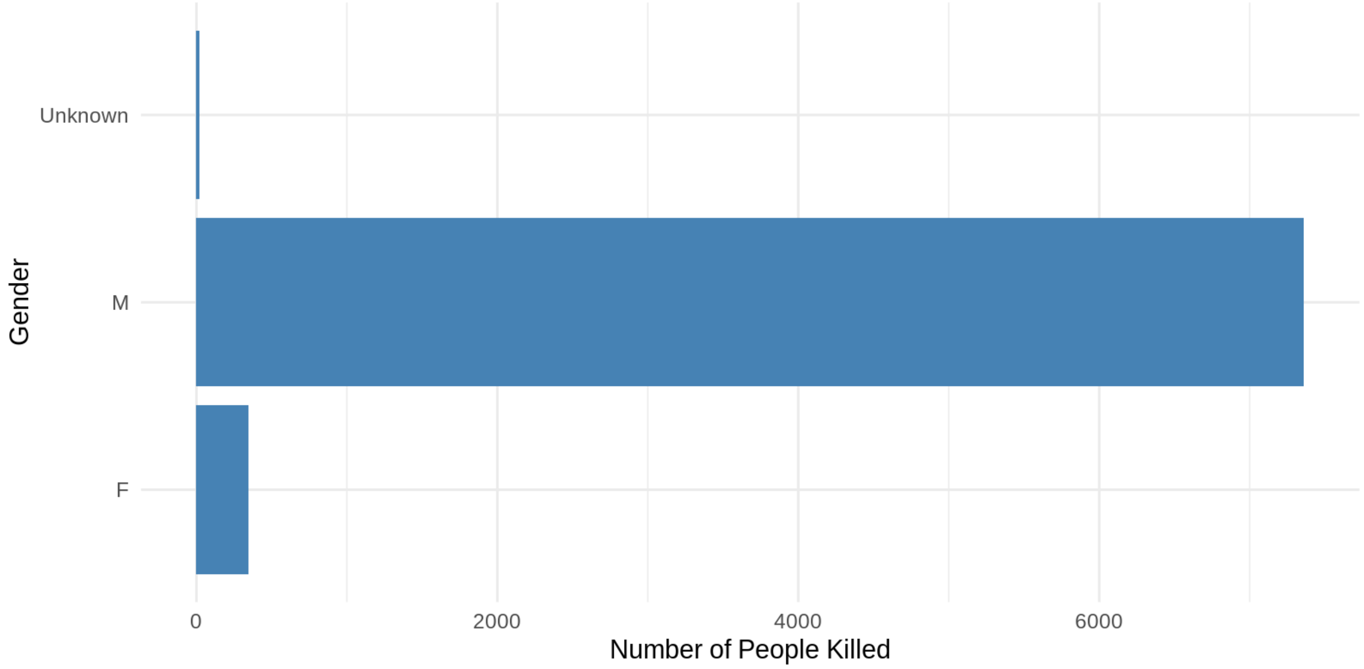
In [ ]:

```
df <- data.frame(data)
```

> Here I have stored my data into a new data frame called `df`

In [329…

```
p1 <- ggplot(df, aes(y=gender))+
  geom_bar(fill="steelblue")+
  labs(x="Number of People Killed",y="Gender")+
  theme_minimal()
p1
```

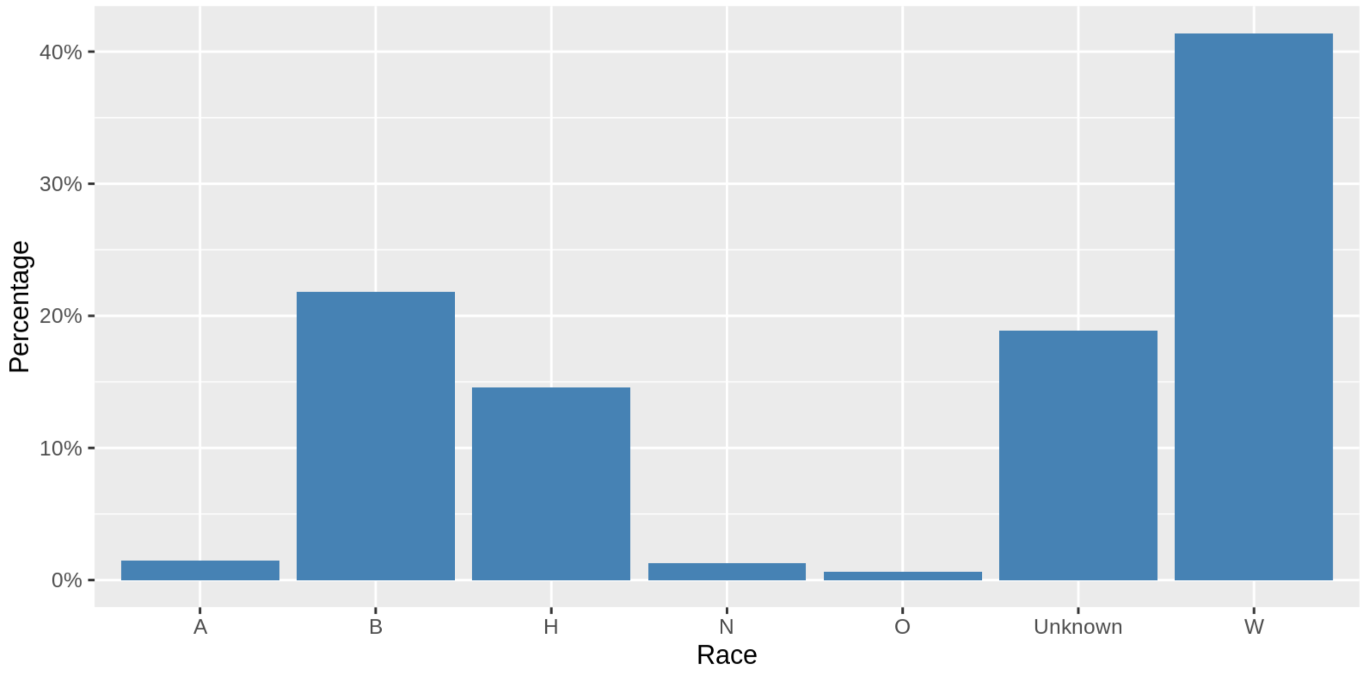


If we look at the bar plot above then we can clearly understand that Number of male victims are way higher than female victims

2. People killed based on their Race .

In [ ]:

```
p2 <- ggplot(df, aes(race)) +
          geom_bar(aes(y = (..count..)/sum(..count..)), fill="steelblue") +
          scale_y_continuous(labels=scales::percent) +
  labs(x="Race",y="Percentage")
p2
```



The above bar plot suggests that:

- Race is unknown of around 18-19% people.
- White,Non Hispanic people killed account for around 41%
- Black People acccount for 21-23%
- Hispanic people account for less than 15%
- Asian people killed account for 1% similar to Native Americans.

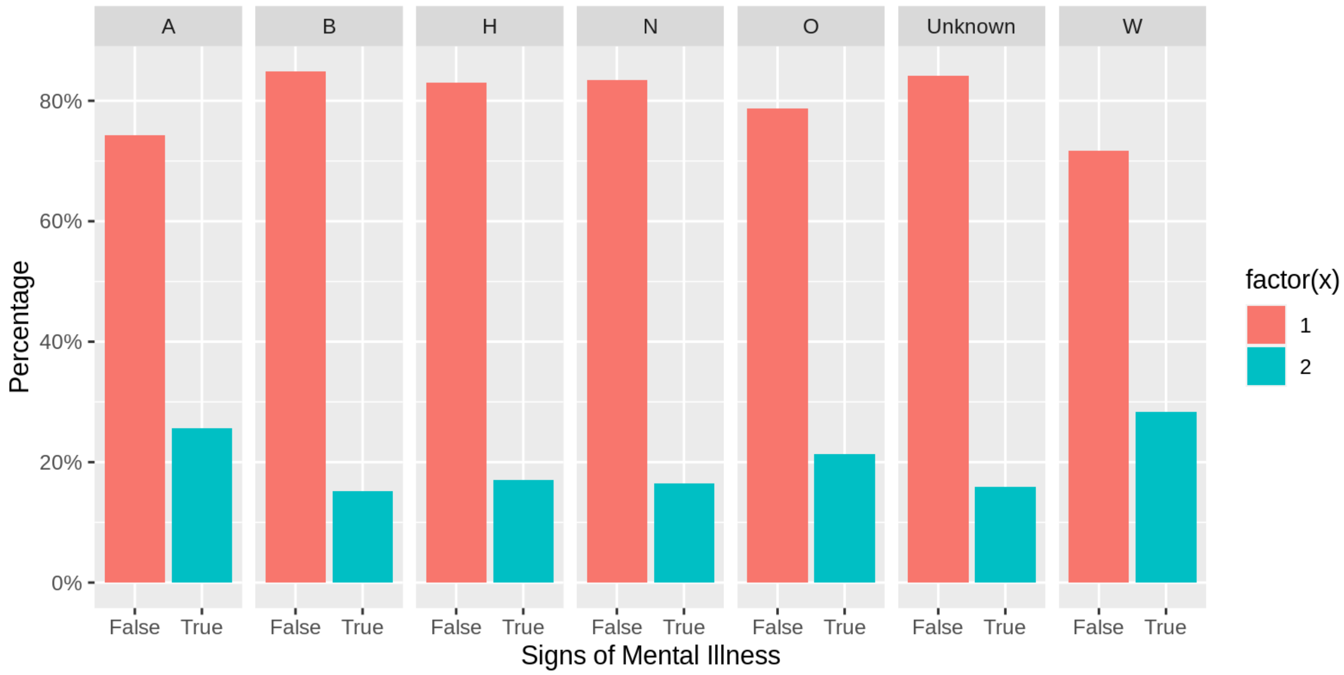
3. People who showed Signs of Mental illness with respect to their race.

- True : Had Mental Illness.
- False : No Mental Illness Present.

In [ ]:

```
p3 <- ggplot(df, aes(signs_of_mental_illness, group = race)) +
          geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
          scale_y_continuous(labels=scales::percent) +
          labs(x="Signs of Mental Illness",y="Percentage") +
          facet_grid(~race)

p3
```



If we observe the above graph, we can state that Out of all the people killed, Mental Illness found in White and Asians is almost same, just around 22-24%.

Victims who were Black, Hispanic and Native Americans had mental illness just around 17-18%

Creating Age Group for all the Victims

In [ ]:

```
df$AgeGroup <- cut(df$age,breaks = c(-Inf ,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85, I

                     labels=c("0-4 years","5-9 years","10-14 years","15-19 years","20-24 y
                     "50-54 years","55-59 years","60-64 years","65-69 years","70-74 years'
                     ,

                     right = FALSE)
```

In [ ]:

```
head(df)
```

A data.frame: 6 × 18

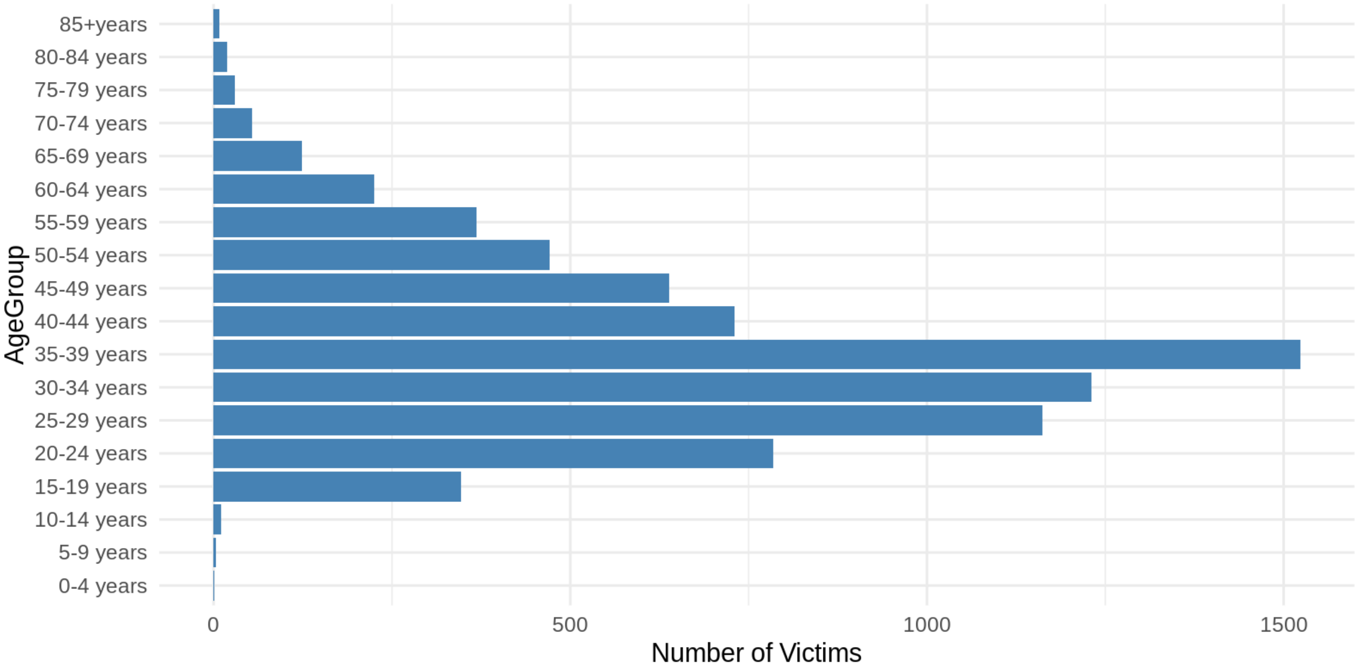
	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_ment
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	
1	3	Tim Elliot	2015-01-02	shot	gun	53	M	A	Shelton	WA	
2	4	Lewis Lee Lembke	2015-01-02	shot	gun	47	M	W	Aloha	OR	
3	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23	M	H	Wichita	KS	
4	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32	M	W	San Francisco	CA	
5	9	Michael Rodriguez	2015-01-04	shot	nail gun	39	M	H	Evans	CO	

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_ment
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	
6	11	Kenneth Joe Brown	2015-01-04	shot	gun	18	M	W	Guthrie	OK	

4. Victims According to their Age group

In [ ]:

```
p4 <- ggplot(df, aes(y = AgeGroup))+
  geom_bar(fill="steelblue")+
  labs(x="Number of Victims") +
  theme_minimal()
p4
```



Victims who got killed were mostly from the age group of 20 years to 44 years.

The highest number of victims were from the age group of 35 - 39 Years.

Maximum Age of a Victim

In [ ]:

```
max(df$age)
```

92

Minimum Age of a Victim

In [ ]:

```
min(df$age)
```

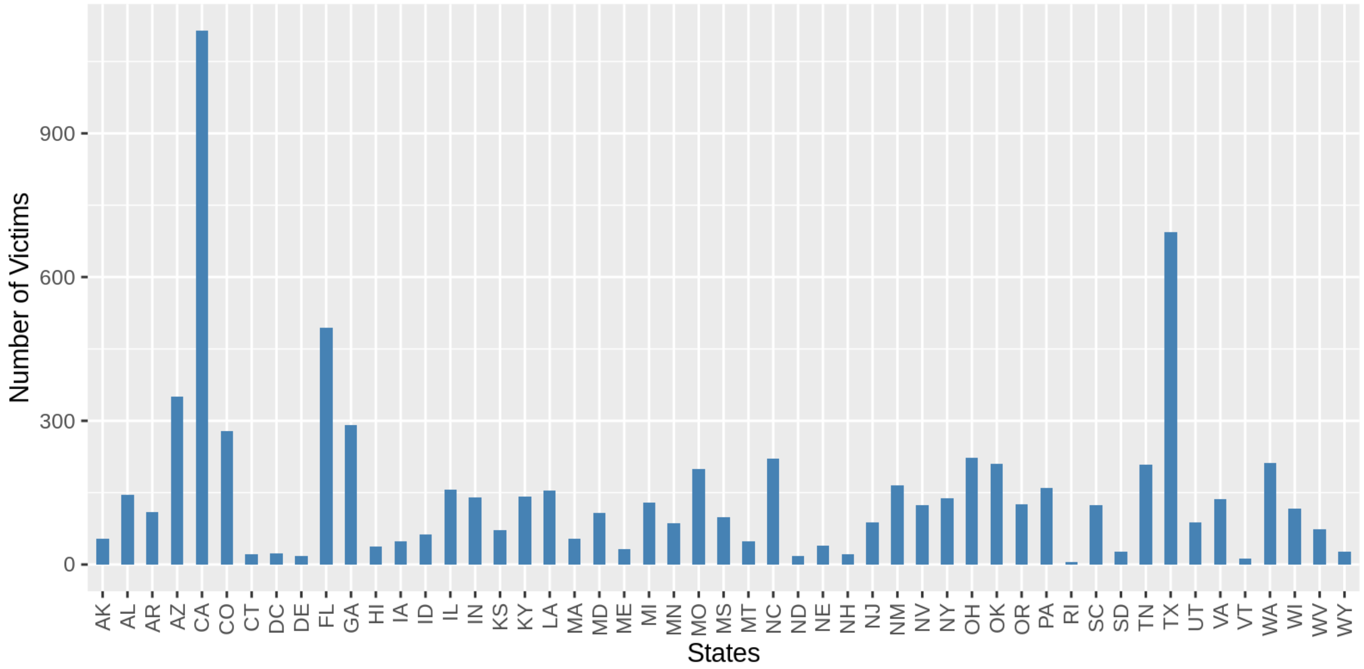
2

5. Victims Killed by State

In [ ]:

```
p5 <- ggplot(df, aes(x = state)) +
      geom_bar(fill="steelblue", width = 0.50) +
      labs(x="States",y="Number of Victims") +
      theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

p5
```



Weapons used by People who were Killed

In [ ]:

```
unique(df$armed)
```

```
'gun' · 'unarmed' · 'toy weapon' · 'nail gun' · 'knife' · 'Unknown' · 'shovel' · 'vehicle' · 'hammer' · 'hatchet' ·
'sword' · 'machete' · 'box cutter' · 'undetermined' · 'metal object' · 'screwdriver' · 'lawn mower blade' ·
'flagpole' · 'guns and explosives' · 'cordless drill' · 'crossbow' · 'BB gun' · 'metal pole' · 'Taser' · 'metal pipe' ·
'metal hand tool' · 'blunt object' · 'metal stick' · 'sharp object' · 'meat cleaver' · 'carjack' · 'chain' ·
'contractor\'s level' · 'railroad spikes' · 'stapler' · 'beer bottle' · 'unknown weapon' · 'binoculars' · 'pellet gun' ·
'bean-bag gun' · 'baseball bat and fireplace poker' · 'straight edge razor' · 'gun and knife' · 'ax' · 'brick' ·
'baseball bat' · 'hand torch' · 'chain saw' · 'garden tool' · 'pair of scissors' · 'pole' · 'pick-axe' · 'flashlight' ·
'baton' · 'spear' · 'chair' · 'pitchfork' · 'hatchet and gun' · 'rock' · 'piece of wood' · 'pipe' · 'glass shard' ·
'motorcycle' · 'pepper spray' · 'metal rake' · 'crowbar' · 'oar' · 'machete and gun' · 'tire iron' · 'air conditioner' ·
'pole and knife' · 'baseball bat and bottle' · 'fireworks' · 'pen' · 'chainsaw' · 'gun and sword' · 'gun and car' ·
'claimed to be armed' · 'incendiary device' · 'samurai sword' · 'bow and arrow' · 'gun and vehicle' ·
'vehicle and gun' · 'wrench' · 'walking stick' · 'barstool' · 'BB gun and vehicle' · 'wasp spray' · 'air pistol' ·
'Airsoft pistol' · 'baseball bat and knife' · 'vehicle and machete' · 'ice pick' · 'car, knife and mace' · 'bottle' ·
'gun and machete' · 'microphone' · 'knife and vehicle' · 'machete and hammer' · 'stake' ·
'incendiary weapon' · 'ax and machete' · 'hammer and garden tool' · 'flare gun' ·
'knife, hammer and gasoline can'
```

Count of Unarmed people

In [ ]:

```
sum(df$armed == 'unarmed')
```

```
456
```

In [ ]:

```
sum(df$armed == 'Unknown')
```

```
209
```

Count of People with Gun

In [ ]:

```
sum(df$armed == 'gun')
```

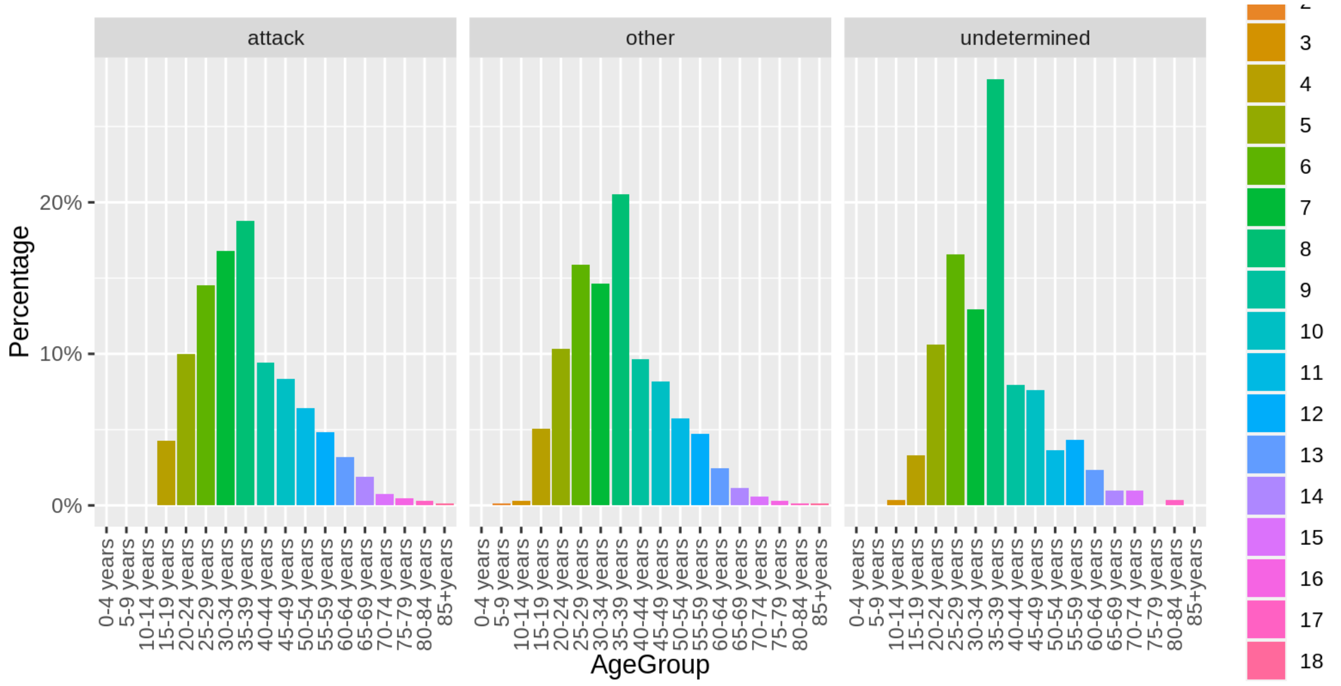
```
4407
```

6. Threat Level on the basis of Race

In [ ]:

```
p6 <- ggplot(df, aes(AgeGroup, group = threat_level)) +
        geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
        scale_y_continuous(labels=scales::percent) +
        ylab("Percentage") +
        theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
        facet_grid(~threat_level)

p6
```



7. Police Shootings based on Year

In [ ]:

```
df$year <- format(as.Date(df$date, format="%Y-%m-%d"),"%Y")
```

In [ ]:

```
head(df,3)
```

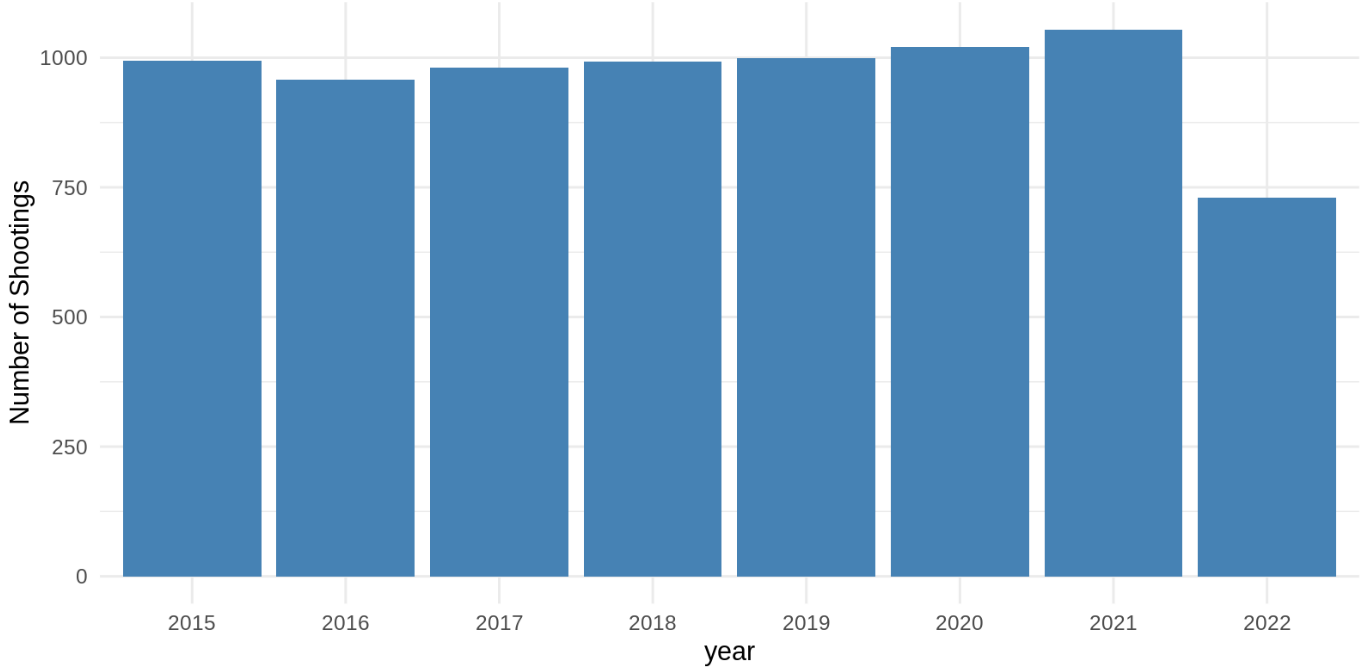
A data.frame: 3 ×

id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	
1	3	Tim Elliot	2015-01-02	shot	gun	53	M	A	Shelton	WA	
2	4	Lewis Lee Lembke	2015-01-02	shot	gun	47	M	W	Aloha	OR	
3	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23	M	H	Wichita	KS	

In [ ]:

```
p7 <- ggplot(df, aes(x=year,group=year)) +
      geom_bar(fill='steelblue') +
      labs(y="Number of Shootings") +
      theme_minimal()
p7
```



In [ ]:

Question 1:

Which of the following techniques might be useful in addressing questions arising from this data set?

```
classification
regression
cluster analysis
```

Now let's discuss the technique which will be most

suitable for this Washinton Post Police Shooting Dataset.

To come to a conclusion we must first understand how Linear Regression, Classification and Clustering work and It will eventually help us decide which technique is best suited in the Scenario.

Regression

- **Linear Regression** - In Linear Regression we try to predict the value of a variable based on the other variables available in the dataset. The output variable which we are predicting is known as Dependent variable and Input variables are known as independent variables.

- The most important thing to remember while working on Linear Regression is that the output is always **Quantitative** and is based on **input-output** observation, without a proper input-output observation we cannot use Linear Regression.

- Now, if we observe our data closely then we can clearly understand that most of the variables available to us have a **Character Based varibale** and only **Age** is a **Integer Based Varibale** and **Longitude and Latitude** are **Numerical Based variable**.

- So suppose if we try to apply linear regression on this data, we won't be able to do it, because we don't have much variable available to use in a model and get results.

- ## Hence we can state that Regression will not be usefull to us in this Dataset.

In [ ]:

```
str(data)
```

```
'data.frame':	7729 obs. of  17 variables:
 $ id                     : int  3 4 5 8 9 11 13 15 16 17 ...
 $ name                   : chr  "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew
Hoffman" ...
 $ date                   : chr  "2015-01-02" "2015-01-02" "2015-01-03" "2015-01-04" ...
 $ manner_of_death        : chr  "shot" "shot" "shot and Tasered" "shot" ...
 $ armed                  : chr  "gun" "gun" "unarmed" "toy weapon" ...
 $ age                    : num  53 47 23 32 39 18 22 35 34 47 ...
 $ gender                 : chr  "M" "M" "M" "M" ...
 $ race                   : chr  "A" "W" "H" "W" ...
 $ city                   : chr  "Shelton" "Aloha" "Wichita" "San Francisco" ...
 $ state                  : chr  "WA" "OR" "KS" "CA" ...
 $ signs_of_mental_illness: chr  "True" "False" "False" "True" ...
 $ threat_level           : chr  "attack" "attack" "other" "attack" ...
 $ flee                   : chr  "Not fleeing" "Not fleeing" "Not fleeing" "Not fleeing" ...
 $ body_camera            : chr  "False" "False" "False" "False" ...
 $ longitude              : chr  "-123.122" "-122.892" "-97.281" "-122.422" ...
 $ latitude               : chr  "47.247" "45.487" "37.695" "37.763" ...
 $ is_geocoding_exact     : chr  "True" "True" "True" "True" ...
```

In [ ]:

Classification

- In Classfication we try to predict the class/categories in terms of Qualitative Output eg: Sick/Not Sick. In classfication we have specific set of inputs given to us which help us decide the category of our data. For example - The Famous Iris dataset which contains 4 Features(Length and Width of Sepals and Petals) helps us classify the Into three categories Iris Setosa, Iris Virginica and Iris Versicolor.

- Now if we look at our current `Fatal_police_shootings_data.csv` we already have few classes available to us like `manner_of_death` , `signs_of_mental_illness` , `threat_level` , `flee` but we have very less information related to **features**.

- Considering the current available variables, we can still apply Classification and try to categorize our data into few classes.

Class `Dangerous/Not Dangerous` - Based on the available variables such as `threat_level` , `Armed` and `Flee` , We can classify person shot into two classes **Dangerous** and **Not Dangerous**.

- `threat_level` - has two categories **Attack** and **Undetermined/Other** where **Attack** being most dangerous as Victim tried to shot/attack police officer or person near to them, and all the different scenarios fall in **Undertermined and Others**.
- `Armed` - This column tells us whether the person had some kind of weapon or was Unarmed.
- `Flee` - This column tells us whether a person was trying to **Flee** (Car, foot) or was **Not Fleeing**.
- If a Person was **UNARMED** , had `attack` column defined as **OTHER/UNDETERMINED** and Was not **NOT FLEEING** from the scene. We can define him as **NOT DANGEROUS**.
- If a Person was **ARMED** had `attack` variable defined as **ATTACK** and was **FLEEING** from the scene. We can define him as **DANGEROUS**.

Similary, based on same variables we can create a **Class** where we can try to see if a Poice shooting was **JUSTIFIED/NOT JUSTIFIED**

- If a Person was **UNARMED** , had `attack` column defined as **OTHER/UNDETERMINED** , and Was not **NOT FLEEING** from the scene. We can say **Not Justified** and other means could have taken.

- If a Person was **ARMED**, had `attack` variable defined as **ATTACK**, and was **FLEEING(Car,Foot)** from the scene. We can define the shooting as **JUSTIFIED**

- ### To conclude we can say that we can apply classification on some levels on the current dataset and if we more features are available, classification can give us deep insights about the data.

```
In [ ]:
```

Clustering

- In Cluster Analysis, we create clusters of based on the given data. We can group them based on their similarities within the dataset and also objects which are not similar can be combined into the clusters.

- The possibility of creating clusters is huge as we can group objects based on the available features/variables and then analyze them.

- # In `Fatal_police_shootings_data.csv` we can't use clustering much looking at the features available to us.We can create small groups based on `armed(weapons they were carrying)` , `threat_level(attack)` , `flee(Fleeing from authority)` . We can also further divide them State-Wise to understand better, Which state had most number Gun Owners and were potential criminals in terms of `threat_level` , `armed` , `flee` .
- # But to perform clustering on a much deeper level, I think presence of features is very important. Most of the features available here can mostly be used to perfrom classification and also most of the answers can be answered via Exploratory Data Analysis.

In [ ]:

```
```

Valuable Resources which helped me with R Programming

https://www.youtube.com/watch?v=qE_nQPojhhw

https://www.youtube.com/watch?v=j1cvFXak_UU

https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/

http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

https://ggplot2.tidyverse.org/reference/geom_bar.html

https://www.statology.org/r-count-number-of-occurrences-in-column/

https://stackoverflow.com/questions/15629192/making-a-bar-chart-in-ggplot-with-vertical-labels-in-x-axis

https://www.sfu.ca/~mjbrydon/tutorials/BAinR/filter.html

https://www.marsja.se/how-to-extract-year-from-date-in-r-with-examples/

https://www.statology.org/extract-year-from-date-in-r/