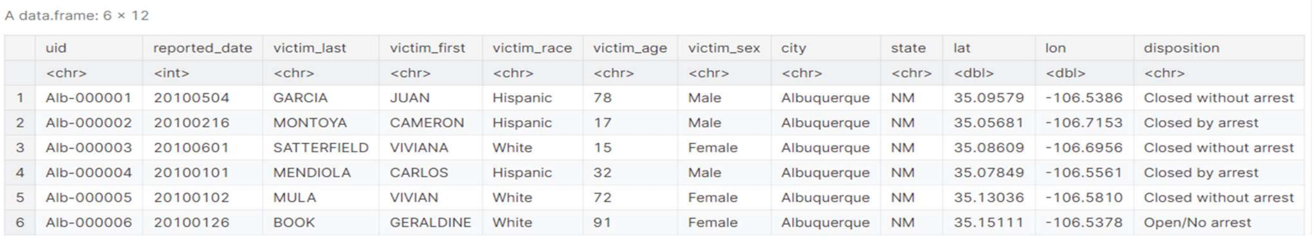
Getting Data set to the Platform

```
df <- read.csv('/kaggle/input/homicides/homicide-data.csv')
head(df)
```



A data.frame: 6 × 12

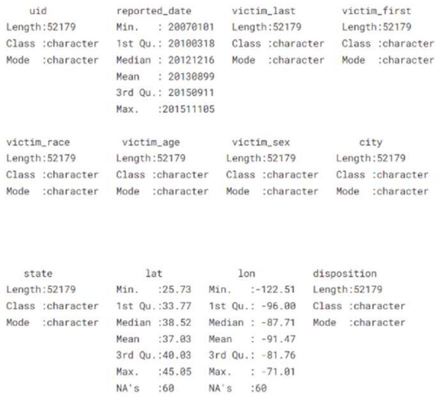
	uid	reported_date	victim_last	victim_first	victim_race	victim_age	victim_sex	city	state	lat	lon	disposition
	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
1	Alb-000001	20100504	GARCIA	JUAN	Hispanic	78	Male	Albuquerque	NM	35.09579	-106.5386	Closed without arrest
2	Alb-000002	20100216	MONTOYA	CAMERON	Hispanic	17	Male	Albuquerque	NM	35.05681	-106.7153	Closed by arrest
3	Alb-000003	20100601	SATTERFIELD	VIVIANA	White	15	Female	Albuquerque	NM	35.08609	-106.6956	Closed without arrest
4	Alb-000004	20100101	MENDIOLA	CARLOS	Hispanic	32	Male	Albuquerque	NM	35.07849	-106.5561	Closed by arrest
5	Alb-000005	20100102	MULA	VIVIAN	White	72	Female	Albuquerque	NM	35.13036	-106.5810	Closed without arrest
6	Alb-000006	20100126	BOOK	GERALDINE	White	91	Female	Albuquerque	NM	35.15111	-106.5378	Open/No arrest

Through this Data set,

- Distribution of the Age and identifying Min, Max, and Quantiles for ages.
- Distribution of ages by victim race and contribution of victims by race.
- To see the contribution of victim's gender by Races.
- To identify which location has reported the most cases and to identify the most cases reported in which year.
- To identify the trend of victim cases over time.
- To identify the most dangerous area.

Investigating and summarizing the data set

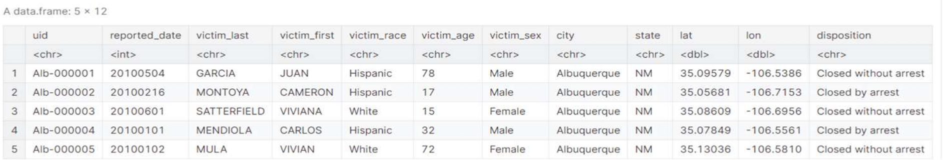
```
summary(df)
```



```
     uid            reported_date       victim_last        victim_first     
 Length:52179       Min.   : 20070101   Length:52179       Length:52179      
 Class :character   1st Qu.: 20100318   Class :character   Class :character  
 Mode  :character   Median : 20121216   Mode  :character   Mode  :character  
                    Mean   : 20130899                                        
                    3rd Qu.: 20150911                                        
                    Max.   :201511105                                        

 victim_race         victim_age         victim_sex            city          
 Length:52179       Length:52179       Length:52179       Length:52179      
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  



    state                lat             lon          disposition       
 Length:52179       Min.   :25.73   Min.   :-122.51   Length:52179      
 Class :character   1st Qu.:33.77   1st Qu.: -96.00   Class :character  
 Mode  :character   Median :38.52   Median : -87.71   Mode  :character  
                    Mean   :37.03   Mean   : -91.47                     
                    3rd Qu.:40.03   3rd Qu.: -81.76                     
                    Max.   :45.05   Max.   : -71.01                     
                    NA's   :60      NA's   :60                          
```

Checking on the first and last 5 Rows in the data set

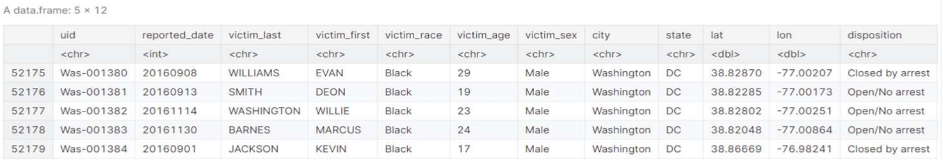
```
cat("First 5 rows in the data Set", "\n")
df[1:5,]
cat("\n")
cat("Last 5 rows in the data set", "\n")
df[-(1:52174),]
```



A data.frame: 5 × 12

	uid	reported_date	victim_last	victim_first	victim_race	victim_age	victim_sex	city	state	lat	lon	disposition
	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
1	Alb-000001	20100504	GARCIA	JUAN	Hispanic	78	Male	Albuquerque	NM	35.09579	-106.5386	Closed without arrest
2	Alb-000002	20100216	MONTOYA	CAMERON	Hispanic	17	Male	Albuquerque	NM	35.05681	-106.7153	Closed by arrest
3	Alb-000003	20100601	SATTERFIELD	VIVIANA	White	15	Female	Albuquerque	NM	35.08609	-106.6956	Closed without arrest
4	Alb-000004	20100101	MENDIOLA	CARLOS	Hispanic	32	Male	Albuquerque	NM	35.07849	-106.5561	Closed by arrest
5	Alb-000005	20100102	MULA	VIVIAN	White	72	Female	Albuquerque	NM	35.13036	-106.5810	Closed without arrest

```
Last 5 rows in the data set
```



A data.frame: 5 × 12

	uid	reported_date	victim_last	victim_first	victim_race	victim_age	victim_sex	city	state	lat	lon	disposition
	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
52175	Was-001380	20160908	WILLIAMS	EVAN	Black	29	Male	Washington	DC	38.82870	-77.00207	Closed by arrest
52176	Was-001381	20160913	SMITH	DEON	Black	19	Male	Washington	DC	38.82285	-77.00173	Open/No arrest
52177	Was-001382	20161114	WASHINGTON	WILLIE	Black	23	Male	Washington	DC	38.82802	-77.00251	Open/No arrest
52178	Was-001383	20161130	BARNES	MARCUS	Black	24	Male	Washington	DC	38.82048	-77.00864	Open/No arrest
52179	Was-001384	20160901	JACKSON	KEVIN	Black	17	Male	Washington	DC	38.86669	-76.98241	Closed by arrest

After going through the Data Set

- Data frame: 52,179 Observations(rows) and 12 Variables(columns)
- Variables: From variables, 9 variables are characters, one is an integer, and the other two variables are numbers (double)
- Missing Values: lat : Latitude variable has 60 missing data(NA). lon : Longitude variable is 60 missing data(NA).
- Data type: reported_date: date format is 20100504 . Need to convert to the appropriate format. Victim_age: Data type is character. Need to convert to the integer.

Checking for unique categories in victim_race, victim_sex, state and disposition

```
df_checking <- df[,c(5,7,9,12)]
colname <- colnames(df_checking)

for(col in colname)
{
    uniques_values <- unique(df[,col])
    unique_counts <- length(unique(df[,col]))
    cat("Unique categories for ",col, "(", unique_counts,")","\n" )
    cat("Unique values are : ",uniques_values,"\n\n")
}
```

```
Unique categories for  victim_race ( 6 )
Unique values are :  Hispanic White Other Black Asian Unknown

Unique categories for  victim_sex ( 3 )
Unique values are :  Male Female Unknown

Unique categories for  state ( 28 )
Unique values are :  NM GA MD LA AL MA NY NC IL OH TX CO MI CA IN FL MO NV KY TN wI MN OK NE PA AZ VA DC

Unique categories for  disposition ( 3 )
Unique values are :  Closed without arrest Closed by arrest Open/No arrest
```

I see here **wI** instead of **WI**

Cleaning the Data Set

```
na_counts <- colSums(is.na(df))
na_counts
```

uid: 0 **reported_date:** 0 **victim_last:** 0 **victim_first:** 0 **victim_race:** 0 **victim_age:** 0 **victim_sex:** 0 **city:** 0 **state:** 0 **lat:** 60 **lon:** 60 **disposition:** 0

We see here only lat and lon has missing values.But, I see reported_date and victim_age . After we change to the correct data type we may receive na values. We can check it . First we need to convert to the data type.

- Here I start with repoted_date - Converting to the Date data type

```
df$reported_date <- ymd(df$reported_date)
# checking  how many na columns has generated after change the data type of the Date
na_counts <- colSums(is.na(df))
na_counts
# Now I see two na missing data under reported_date

# I want see which observations has the problem
problem_rows_rpt_date <- df[is.na(df$reported_date),]
problem_rows_rpt_date
```

uid: 0 **reported_date:** 2 **victim_last:** 0 **victim_first:** 0 **victim_race:** 0 **victim_age:** 2999 **victim_sex:** 0 **city:** 0 **state:** 0 **lat:** 60 **lon:** 60 **disposition:** 0

A data.frame: 2 × 12

	uid	reported_date	victim_last	victim_first	victim_race	victim_age	victim_sex	city	state	lat	lon	disposition
	<chr>	<date>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
33585	Mia-000649	NA	SALAS	LUIS	Hispanic	NA	Male	Miami	FL	25.76990	-80.21719	Closed by arrest
33588	Mia-000652	NA	BUNCH	GERALD A.	Black	NA	Male	Miami	FL	25.82695	-80.20212	Open/No arrest

- Move on to the victim_age

```
df$victim_age <- as.numeric(df$victim_age, na.rm = TRUE)
na_counts <- colSums(is.na(df))
na_counts
```

uid: 0 **reported_date:** 2 **victim_last:** 0 **victim_first:** 0 **victim_race:** 0 **victim_age:** 2999 **victim_sex:** 0 **city:** 0 **state:** 0 **lat:** 60 **lon:** 60 **disposition:** 0

```
#Cheking a few rows from the data set which has the missing data for victim_age
problem_rows_victim_age<- df[is.na(df$victim_age),]
head(problem_rows_victim_age)
```

Now I see 2999 missing values in the victim_age's variable after convert to the number.

A data.frame: 6 × 12

	uid	reported_date	victim_last	victim_first	victim_race	victim_age	victim_sex	city	state	lat	lon	disposition
	<chr>	<date>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
12	Alb-000012	2010-02-18	LUJAN	KEVIN	White	NA	Male	Albuquerque	NM	35.07701	-106.5649	Closed without arrest
60	Alb-000060	2011-05-30	ORTIZ-BURCIAGA	VICTORIA	Hispanic	NA	Female	Albuquerque	NM	NA	NA	Open/No arrest
103	Alb-000103	2012-04-28	VALERIO	MAY	Unknown	NA	Unknown	Albuquerque	NM	35.08802	-106.5631	Closed by arrest
122	Alb-000122	2012-10-26	MACAIO	WESTFALL	White	NA	Female	Albuquerque	NM	35.13288	-106.5263	Closed by arrest
165	Alb-000165	2014-02-08	MONTANO	IZABELLAH	Other	NA	Female	Albuquerque	NM	35.07912	-106.5139	Closed by arrest
186	Alb-000186	2016-02-05	PURVIS	GEORGE JR.	Unknown	NA	Unknown	Albuquerque	NM	35.07343	-106.5487	Closed by arrest

Now I'm going to drop all **"na"** from the data set

```
df2 <- na.omit(df)
# cheking again are there any missing values under variables

na_counts <- colSums(is.na(df2))
na_counts
```

uid: 0 **reported_date:** 0 **victim_last:** 0 **victim_first:** 0 **victim_race:** 0 **victim_age:** 0 **victim_sex:** 0 **city:** 0 **state:** 0 **lat:** 0 **lon:** 0 **disposition:** 0

No missing vaules under the variables now

I'm going to change wI to WI

```
df2$state <- gsub("wI","WI",df2$state)

# Checking unique values under State variable to make sure Data has been changed.
unique(df2[,9])
```

'NM' 'GA' 'MD' 'LA' 'AL' 'MA' 'NY' 'NC' 'IL' 'OH' 'CO' 'MI' 'TX' 'CA' 'IN' 'FL' 'MO' 'NV' 'KY' 'TN' 'WI' 'MN' 'OK' 'NE' 'PA' 'VA' 'DC'

Now It has been changed

Analysis and Visualization

Identifying Outliers

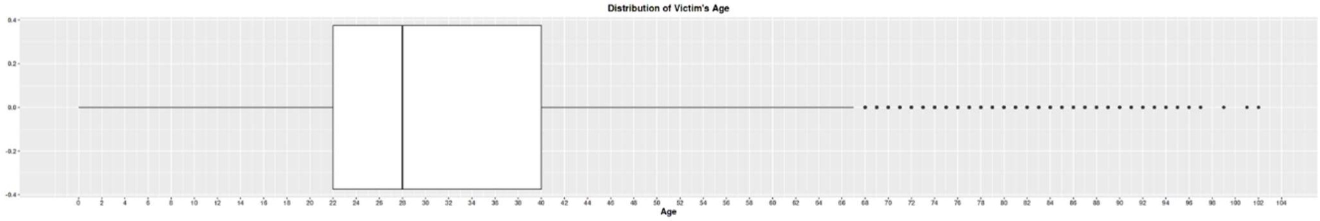
```
summary(df2)
```

```
     uid             reported_date         victim_last         victim_first
 Length:49122       Min.   :2007-01-01   Length:49122       Length:49122
 Class :character   1st Qu.:2010-03-19   Class :character   Class :character
 Mode  :character   Median :2012-12-16   Mode  :character   Mode  :character
                    Mean   :2012-11-03
                    3rd Qu.:2015-09-13
                    Max.   :2017-12-31
 victim_race          victim_age     victim_sex             city
 Length:49122       Min.   :  0.0   Length:49122       Length:49122
 Class :character   1st Qu.: 22.0   Class :character   Class :character
 Mode  :character   Median : 28.0   Mode  :character   Mode  :character
                    Mean   : 31.8
                    3rd Qu.: 40.0
                    Max.   :102.0
    state                lat             lon           disposition
 Length:49122       Min.   :25.73   Min.   :-122.51   Length:49122
 Class :character   1st Qu.:34.04   1st Qu.: -95.47   Class :character
 Mode  :character   Median :38.67   Median : -87.67   Mode  :character
                    Mean   :37.25   Mean   : -90.91
                    3rd Qu.:40.39   3rd Qu.: -81.66
                    Max.   :45.05   Max.   : -71.01
```

I see min of victim_age is 0 and max of victim_age is 102

```
options(repr.plot.width = 30,repr.plot.height = 5 )
```

```
ggplot(df2, aes(x = victim_age))+
scale_x_continuous(breaks = seq(0,105,2))+
geom_boxplot()+labs(title = "Distribution of Victim's Age", x = 'Age')+
theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 15),
     axis.title.x = element_text(face = "bold", size = 14),
     axis.title.y = element_text(face = "bold", size = 14),
     axis.text.x = element_text(face = "bold", size = 10),
     axis.text.y = element_text(face = "bold", size = 10))
```



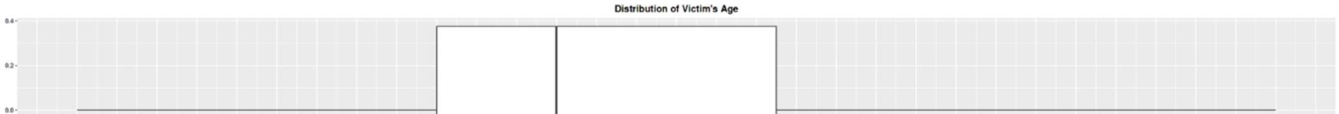
I assumed here that victim_age is less than 3 and more than 64 people don't have physical capability to do the victim. Then I'm going to remove these age range people from the data set.

```
data_set<- df2 %>% filter(victim_age <= 64 & victim_age >3)
ana_data_set <- data.frame(data_set)
summary(ana_data_set)
```

```
     uid            reported_date         victim_last        victim_first
 Length:46649       Min.   :2007-01-01   Length:46649       Length:46649
 Class :character   1st Qu.:2010-03-22   Class :character   Class :character
 Mode  :character   Median :2012-12-22   Mode  :character   Mode  :character
                    Mean   :2012-11-06
                    3rd Qu.:2015-09-15
                    Max.   :2017-12-31
 victim_race          victim_age    victim_sex            city
 Length:46649       Min.   : 4.0   Length:46649       Length:46649
 Class :character   1st Qu.:22.0   Class :character   Class :character
 Mode  :character   Median :28.0   Mode  :character   Mode  :character
                    Mean   :31.2
                    3rd Qu.:39.0
                    Max.   :64.0
    state                lat             lon          disposition
 Length:46649       Min.   :25.73   Min.   :-122.51   Length:46649
 Class :character   1st Qu.:34.05   1st Qu.: -95.44   Class :character
 Mode  :character   Median :38.69   Median : -87.66   Mode  :character
                    Mean   :37.28   Mean   : -90.83
                    3rd Qu.:40.44   3rd Qu.: -81.66
                    Max.   :45.05   Max.   : -71.01
```

After Remove Outliers

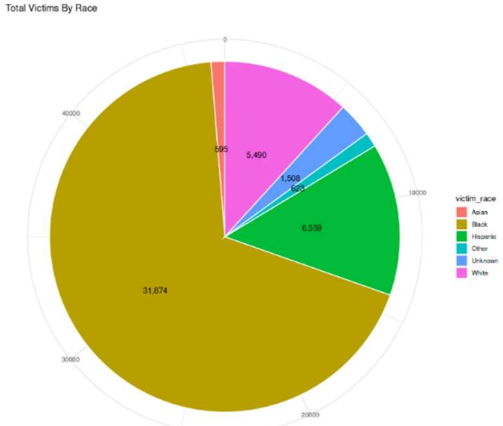
```
ggplot(data_set, aes(x=victim_age)) +
geom_boxplot()+
scale_x_continuous(breaks = seq(0,80,2))+
labs(title = "Distribution of Victim's Age", x = 'Age')+
theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 15),
     axis.title.x = element_text(face = "bold", size = 14),
     axis.title.y = element_text(face = "bold", size = 14),
     axis.text.x = element_text(face = "bold", size = 10),
     axis.text.y = element_text(face = "bold", size = 10))
```



Checking for victim_race and contribution

```
options(repr.plot.width = 10, repr.plot.height = 10)
```

```
data_set %>%
select(victim_race) %>%
group_by(victim_race)%>%
summarise(total_count = (count = n()))%>%
ggplot(aes(x="", y = total_count, fill = victim_race ))+
geom_bar(stat = "identity",width = 6, color = 'white')+
coord_polar("y", start =0)+
geom_text(aes(label =scales::comma(total_count)), position = position_stack(vjust = 0.5))+
labs(title = "Total Victims By Race", x = NULL, y = NULL)+
theme_minimal()+
theme(axis.text.y = element_text(face = "bold", size =15))
```

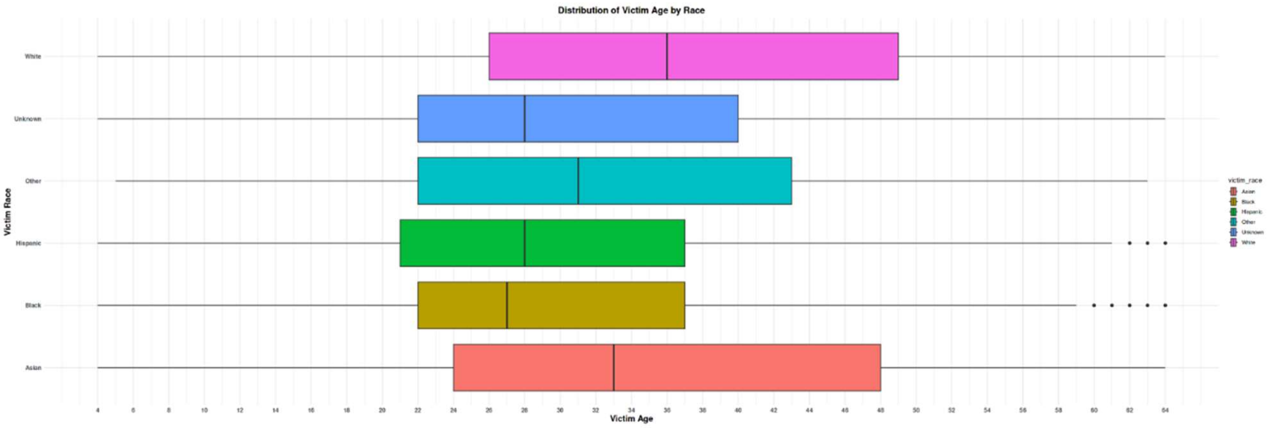


Black people have been involved to victims that other races.

Visualizing Victim Race VS Victim Age in Box plot chart to check the Age Distribution of all Races.

```
options(repr.plot.width = 30)
```

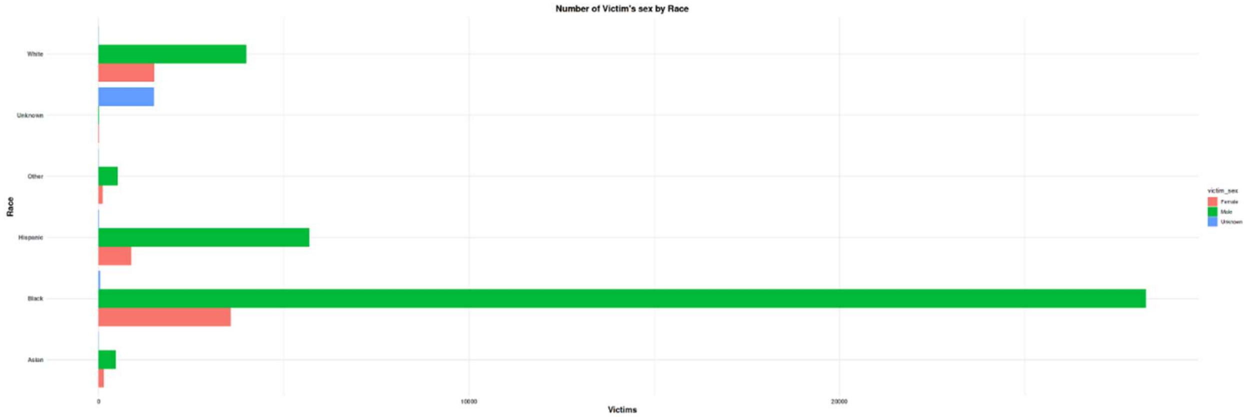
```
data_set %>%
ggplot(aes(x = victim_age, y = victim_race, fill = victim_race))+
geom_boxplot()+
scale_x_continuous(breaks =seq(4,64,2))+
labs(title = "Distribution of Victim Age by Race",  x = "Victim Age", y = "Victim Race")+
theme_minimal()+
theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 15),
     axis.title.x = element_text(face = "bold", size = 14),
     axis.title.y = element_text(face = "bold", size = 14),
     axis.text.x = element_text(face = "bold", size = 10),
     axis.text.y = element_text(face = "bold", size = 10))
```



- White people have the Highest median Age.
- The lowest median age for Black race people.
- Asian People and White people have a wider spread of the age.
- Black people's ages clustered around 22 - 37 age.

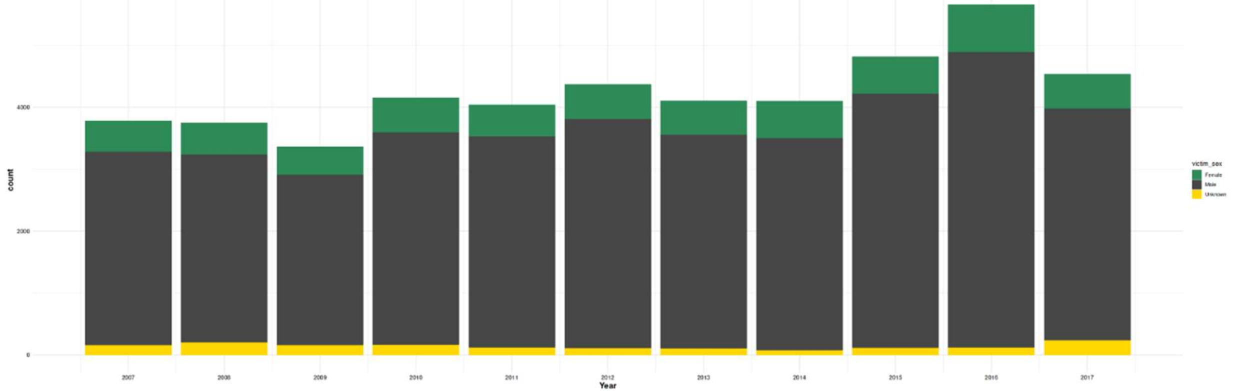
Checking how many Females and males have been involved in victims from all races.

```
data_set %>%
  ggplot(aes(y = victim_race, fill = victim_sex)) +
  geom_bar(stat = "count", position = "dodge") +
  labs(title = "Number of Victim's sex by Race", y = "Race", x = "Victims")+
theme_minimal()+
theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 15),
     axis.title.x = element_text(face = "bold", size = 14),
     axis.title.y = element_text(face = "bold", size = 14),
     axis.text.x = element_text(face = "bold", size = 10),
     axis.text.y = element_text(face = "bold", size = 10))
```



- Black Male has involved in victims than another Race's Male
- Black Females have involved victims than another Race's Female

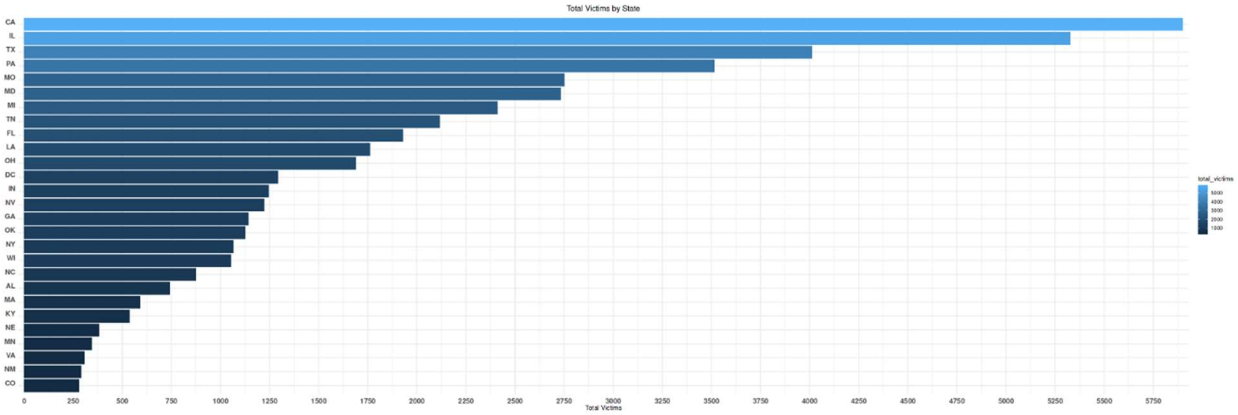
```
Year <- year(data_set$reported_date)

data_set %>%
ggplot(aes(x = Year, fill =victim_sex ))+
scale_x_continuous(breaks = seq(2007,2017,1))+
scale_fill_manual(values = c("Male" = "gray28", "Female" = "seagreen", "Unknown" = "gold")) +
geom_bar()+
theme_minimal()+
theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 15),
     axis.title.x = element_text(face = "bold", size = 14),
     axis.title.y = element_text(face = "bold", size = 14),
     axis.text.x = element_text(face = "bold", size = 10),
     axis.text.y = element_text(face = "bold", size = 10))
```



- I see in 2016 most cases and in 2009 reported less than the other years.

```
options(repr.plot.height = 10)
```

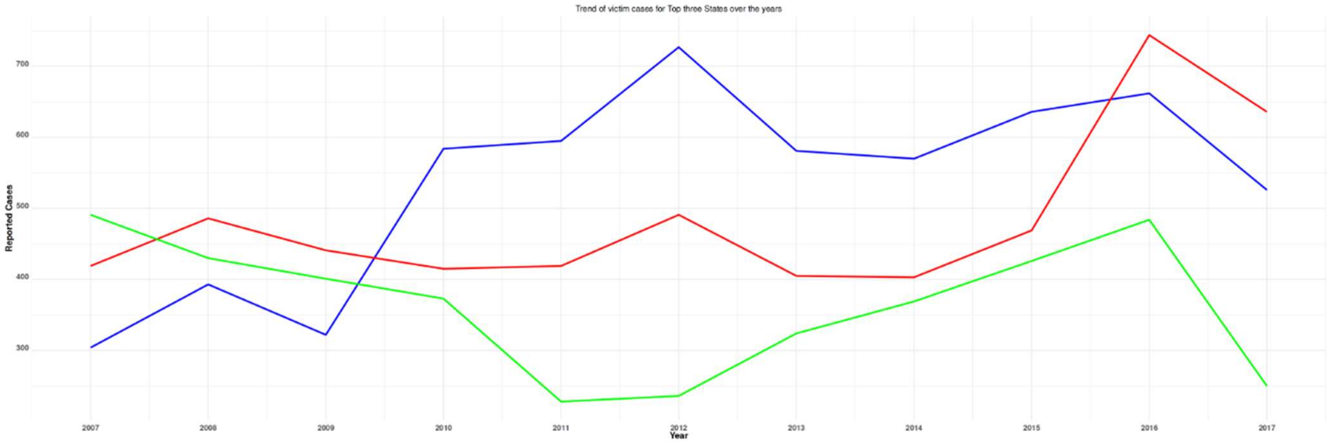
```
data_set %>%
  select(state) %>%
  group_by(state) %>%
  summarise(total_victims = n(), .groups = "drop") %>%
  arrange(desc(total_victims)) %>%
  ggplot(aes(y = reorder(state, total_victims), x = total_victims, fill = total_victims)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(expand =c(0.005,0), breaks = seq(0,6000,250))+
  labs(title = "Total Victims by State",
       x = "Total Victims",
       y = "State") +
  theme_minimal() +
  theme(
      plot.title = element_text(hjust = 0.5),
        axis.title.y = element_blank(),
        axis.text.y = element_text(vjust = 0, face = 'bold', size = 12),
        axis.text.x = element_text(vjust = 0, face = 'bold', size = 12))
```



- Most victims have reported in CA state.
- After CA, next two leading states for victims are IL and TX
- Least cases have been reported in CO - Colorado

I'm going to investigate deeply the data for the states of CA, TX and IL.

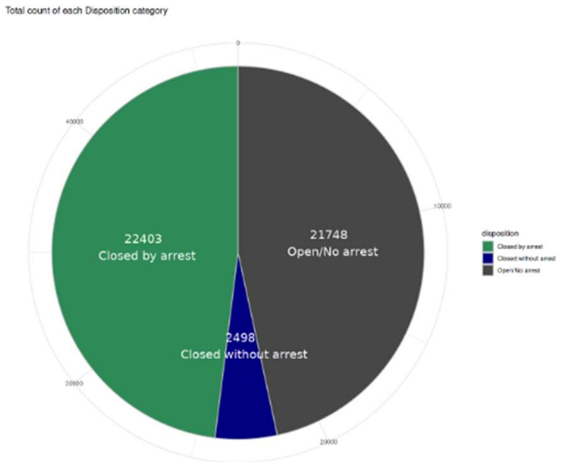
```
data_set %>%
    mutate(year) %>%
    select(year, state)%>%
    group_by(state,year)%>%
    summarise(total_incident = n(), .groups = 'drop')%>%
    filter(state %in% c("CA", "TX", "IL"))%>%
ggplot(aes(x= year, y = total_incident,color = state)) +
  geom_line(linewidth =1) +
  theme_minimal() +
  scale_x_continuous(breaks = seq(2007,2017,1))+
  labs(title = "Trend of victim cases for Top three States over the years",
       x = "Year",
       y = "Reported Cases",
       color = "State") +
  scale_color_manual(values = c("CA" = "blue", "TX" = "green", "IL" = "red"))+
  theme(
      plot.title = element_text(hjust = 0.5),
      axis.title.y = element_text(face = 'bold',size =14),
      axis.title.x = element_text(face = 'bold',size =14),
        axis.text.y = element_text(vjust = 0, face = 'bold', size = 12),
        axis.text.x = element_text(vjust = 0, face = 'bold', size = 12),
  legend.position = 'none')
```



Checking on the Status of the Disposition

- I'm going to create a Pie chart to display this variable. This would help to see the contribution.

```
data_set%>%
    select(disposition)%>%
    group_by(disposition)%>%
    summarise(total_disposition = n())%>%
ggplot(aes(x =" ", y =total_disposition, fill =disposition ))+
    geom_bar(stat = "identity", width = 6, color = "grey")+
    coord_polar("y", start =0)+
    scale_fill_manual(values = c("Closed without arrest" = "navy","Closed by arrest" = "seagreen","Open/No arrest" = "gray28"
))+
    geom_text(aes(label =paste(total_disposition, "\n",disposition)), position = position_stack(vjust =0.5), family = "bold",
color = 'white', size = 6)+
    labs(title = "Total count of each Disposition category", x =NULL, y = NULL)+
    theme_minimal()
```



Still remaining more cases to solve.

Through the Analysis

- I analyzed homicide data which was generated for the period 2007 - 2017

- The maximum age of the data set shows as 102 and the minimum shows as 0. - I see some Victim's ages are not possible to do the victim. It may be a typing mistake or wrong data collection. Should have some mental and physical strength to do the victim.
- CA is the worst place according to the data set (most reported cases).
- CO is the calm place according to the data set (least reported cases).
- Victim_races = "black" is involved in most cases according to the data set information.
- I don't see a huge gap through the year for the victims. But I can say most cases were reported in 2016 and the least cases were reported in 2009.
- According to the data set's information 21748 remaining cases are still to be solved or to be arrested.