Project - US Homicide Statistics Analysis

Load Libraries

library(tidyverse)

## -- Attaching packages ------------------------------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts --------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(skimr)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(wesanderson)

Load Dataset

homicide\_stats <- read\_csv("HomicideStats.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## Year = col\_double(),  
## Incident = col\_double(),  
## `Victim Age` = col\_double(),  
## `Perpetrator Age` = col\_double(),  
## `Victim Count` = col\_double(),  
## `Perpetrator Count` = col\_double()  
## )

## See spec(...) for full column specifications.

Summary of Data

skim(homicide\_stats)

Data summary

|  |  |
| --- | --- |
| Name | homicide\_stats |
| Number of rows | 638454 |
| Number of columns | 24 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 18 |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| Record ID | 0 | 1 | 6 | 6 | 0 | 638454 | 0 |
| Agency Code | 0 | 1 | 7 | 7 | 0 | 12003 | 0 |
| Agency Name | 0 | 1 | 3 | 81 | 0 | 9216 | 0 |
| Agency Type | 0 | 1 | 7 | 16 | 0 | 7 | 0 |
| City | 0 | 1 | 3 | 31 | 0 | 1782 | 0 |
| State | 0 | 1 | 4 | 20 | 0 | 51 | 0 |
| Month | 0 | 1 | 3 | 9 | 0 | 12 | 0 |
| Crime Type | 0 | 1 | 22 | 26 | 0 | 2 | 0 |
| Crime Solved | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| Victim Sex | 0 | 1 | 4 | 7 | 0 | 3 | 0 |
| Victim Race | 0 | 1 | 5 | 29 | 0 | 5 | 0 |
| Victim Ethnicity | 0 | 1 | 7 | 12 | 0 | 3 | 0 |
| Perpetrator Sex | 0 | 1 | 4 | 7 | 0 | 3 | 0 |
| Perpetrator Race | 0 | 1 | 5 | 29 | 0 | 5 | 0 |
| Perpetrator Ethnicity | 0 | 1 | 7 | 12 | 0 | 3 | 0 |
| Relationship | 0 | 1 | 3 | 20 | 0 | 28 | 0 |
| Weapon | 0 | 1 | 3 | 13 | 0 | 16 | 0 |
| Record Source | 0 | 1 | 3 | 4 | 0 | 2 | 0 |

**Variable type: numeric**

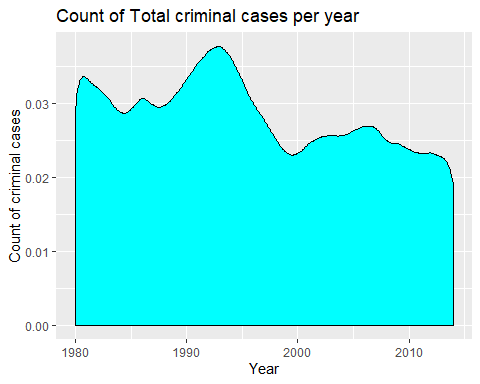
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| Year | 0 | 1 | 1995.80 | 9.93 | 1980 | 1987 | 1995 | 2004 | 2014 | ▇▇▇▆▆ |
| Incident | 0 | 1 | 22.97 | 92.15 | 0 | 1 | 2 | 10 | 999 | ▇▁▁▁▁ |
| Victim Age | 0 | 1 | 35.03 | 41.63 | 0 | 22 | 30 | 42 | 998 | ▇▁▁▁▁ |
| Perpetrator Age | 1 | 1 | 20.32 | 17.89 | 0 | 0 | 21 | 31 | 99 | ▇▇▂▁▁ |
| Victim Count | 0 | 1 | 0.12 | 0.54 | 0 | 0 | 0 | 0 | 10 | ▇▁▁▁▁ |
| Perpetrator Count | 0 | 1 | 0.19 | 0.59 | 0 | 0 | 0 | 0 | 10 | ▇▁▁▁▁ |

summary(homicide\_stats)

## Record ID Agency Code Agency Name Agency Type   
## Length:638454 Length:638454 Length:638454 Length:638454   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## City State Year Month   
## Length:638454 Length:638454 Min. :1980 Length:638454   
## Class :character Class :character 1st Qu.:1987 Class :character   
## Mode :character Mode :character Median :1995 Mode :character   
## Mean :1996   
## 3rd Qu.:2004   
## Max. :2014   
##   
## Incident Crime Type Crime Solved Victim Sex   
## Min. : 0.00 Length:638454 Length:638454 Length:638454   
## 1st Qu.: 1.00 Class :character Class :character Class :character   
## Median : 2.00 Mode :character Mode :character Mode :character   
## Mean : 22.97   
## 3rd Qu.: 10.00   
## Max. :999.00   
##   
## Victim Age Victim Race Victim Ethnicity Perpetrator Sex   
## Min. : 0.00 Length:638454 Length:638454 Length:638454   
## 1st Qu.: 22.00 Class :character Class :character Class :character   
## Median : 30.00 Mode :character Mode :character Mode :character   
## Mean : 35.03   
## 3rd Qu.: 42.00   
## Max. :998.00   
##   
## Perpetrator Age Perpetrator Race Perpetrator Ethnicity Relationship   
## Min. : 0.00 Length:638454 Length:638454 Length:638454   
## 1st Qu.: 0.00 Class :character Class :character Class :character   
## Median :21.00 Mode :character Mode :character Mode :character   
## Mean :20.32   
## 3rd Qu.:31.00   
## Max. :99.00   
## NA's :1   
## Weapon Victim Count Perpetrator Count Record Source   
## Length:638454 Min. : 0.0000 Min. : 0.0000 Length:638454   
## Class :character 1st Qu.: 0.0000 1st Qu.: 0.0000 Class :character   
## Mode :character Median : 0.0000 Median : 0.0000 Mode :character   
## Mean : 0.1233 Mean : 0.1852   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000   
## Max. :10.0000 Max. :10.0000   
##

Univariate Numerical - Crime Rate over the time period

homicide\_stats %>%  
 ggplot(aes(x = Year))+  
 geom\_density(fill = "cyan", color = "black")+  
 labs(title = "Count of Total criminal cases per year",  
 x = "Year",  
 y = "Count of criminal cases")



homicide\_stats %>%  
 group\_by(Year) %>%  
 mutate(Count = n()) %>%  
 select(Year, Count) %>%  
 distinct()

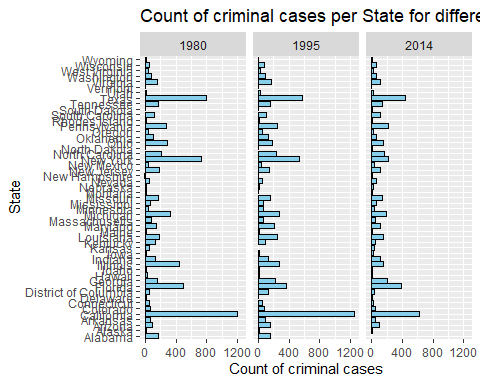
## # A tibble: 35 x 2  
## # Groups: Year [35]  
## Year Count  
## <dbl> <int>  
## 1 1980 23092  
## 2 1981 21208  
## 3 1982 20544  
## 4 1983 19653  
## 5 1984 18093  
## 6 1985 18386  
## 7 1986 20125  
## 8 1987 18783  
## 9 1988 18779  
## 10 1989 19868  
## # ... with 25 more rows

# homicide\_stats %>%  
# ggplot(aes(x = Year, fill = State))+  
# geom\_histogram()+  
# labs(title = "Count of Total criminal cases per year",  
# x = "Year",  
# y = "Count of criminal cases")

Univariate Categorical (three different years - min, median, max)

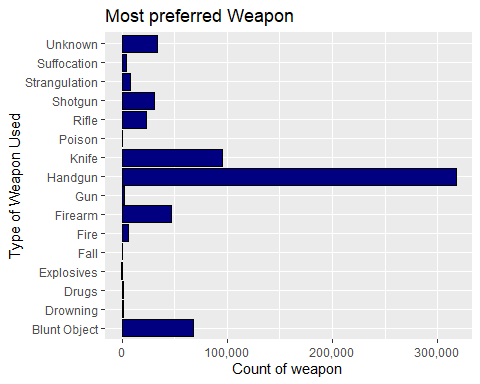
Count of criminal cases per State for different years

homicide\_stats %>%  
 filter(Year == c(min(Year), median(Year), max(Year))) %>%  
 ggplot(aes(x = State))+  
 geom\_bar(fill = "skyblue", color = "black")+  
 facet\_grid( ~ (Year))+  
 coord\_flip()+  
 labs(title = "Count of criminal cases per State for different years",  
 x = "State",  
 y = "Count of criminal cases")



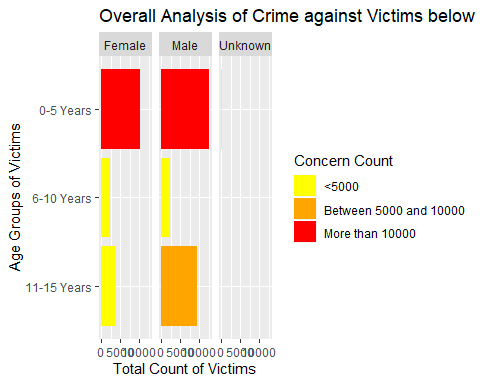
Univariate Categorical - Most Preferred Weapon from 1980-2014

homicide\_stats %>%  
 ggplot(aes(x = Weapon, y = ..count..))+  
 geom\_bar(fill = "navyblue", color = "black")+  
 coord\_flip()+  
 labs(title = "Most preferred Weapon",  
 x = "Type of Weapon Used",  
 y = "Count of weapon ")+  
 scale\_y\_continuous(labels = comma)



Univariate Analysis Numerical - Which age group below 15 years is most victimised?

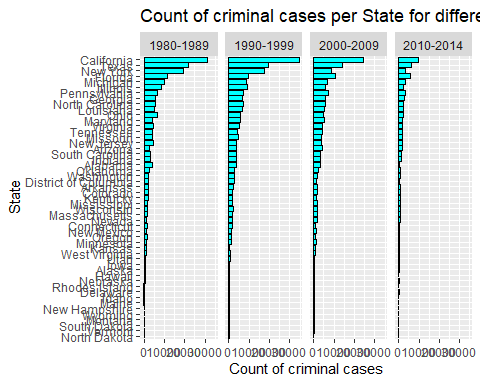
homicide\_stats\_age15 <- homicide\_stats %>%  
 filter(`Victim Age` <= 15) %>%  
 mutate(age\_group = cut(`Victim Age`, breaks=c(0,5,10,15), include.lowest = T, labels = c("0-5 Years", "6-10 Years", "11-15 Years")))  
  
homicide\_stats\_age15 <- homicide\_stats\_age15 %>%  
 group\_by(age\_group, `Victim Sex`) %>%  
 mutate(count = n(),  
 `Concern Count` = cut(count, breaks = c(0,5000,10000,15000), include.lowest = T, labels = c("<5000","Between 5000 and 10000","More than 10000")))  
  
#palette  
pal <- c("<5000" = "yellow",   
 "Between 5000 and 10000" = "orange",   
 "More than 10000" = "red")  
  
homicide\_stats\_age15 %>%  
 filter(!is.na(age\_group)) %>%  
 ggplot(aes(x = age\_group, fill = `Concern Count`))+  
 geom\_bar()+  
 facet\_grid( ~ `Victim Sex`)+  
 labs(title = "Overall Analysis of Crime against Victims below 15 years",  
 x = "Age Groups of Victims",  
 y = "Total Count of Victims")+  
 coord\_flip()+  
 scale\_x\_discrete(limits = rev(levels(homicide\_stats\_age15$age\_group))) +  
 scale\_fill\_manual(  
 values = pal,  
 limits = names(pal))



Univariate Categorical

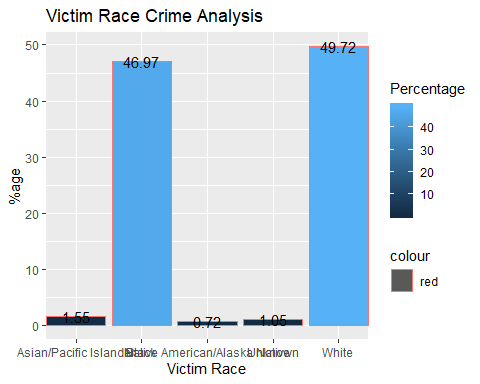
Count of Criminal Cases per State for different decades

#library(forcats)  
homicide\_stats\_decade <- homicide\_stats %>%  
 mutate(decade = cut(Year, breaks = c(1980,1989,1999,2009,2014), include.lowest = T, labels = c("1980-1989","1990-1999","2000-2009","2010-2014")))  
  
homicide\_stats\_decade\_order <- homicide\_stats\_decade %>%  
 select(State, decade) %>%  
 group\_by(State,decade) %>%  
 mutate(Count = n()) %>%  
 distinct()   
  
homicide\_stats\_decade\_order %>%  
 #mutate(Count = fct\_reorder(State,desc(Count))) %>%  
 ggplot(aes(x = reorder(State, Count), y = Count))+  
 geom\_bar(stat = "identity", fill = "cyan", color = "black")+  
 coord\_flip()+  
 facet\_grid( ~ decade)+  
 labs(title = "Count of criminal cases per State for different decades",  
 x = "State",  
 y = "Count of criminal cases")



Victim Race Crime Analysis

homicide\_stats\_victim <- homicide\_stats %>%  
 mutate(Total = n()) %>%  
 group\_by(`Victim Race`) %>%  
 select(`Victim Race`, Total) %>%  
 mutate(Count = n()) %>%  
 mutate(Percentage = round(Count\*100/Total,2)) %>%  
 distinct()  
  
# Gradient color  
#grad <- wes\_palette("Zissou1", 100, type = "continuous")  
  
homicide\_stats\_victim %>%  
 ggplot(aes(x = `Victim Race`, y = Percentage))+  
 geom\_bar(stat = "identity", aes(fill = Percentage, color = "red"))+  
 geom\_text(aes(label = Percentage))+  
 labs(title = "Victim Race Crime Analysis",  
 x = "Victim Race",  
 y = "%age")+   
 scale\_x\_discrete(expand = c(0, 0)) #+

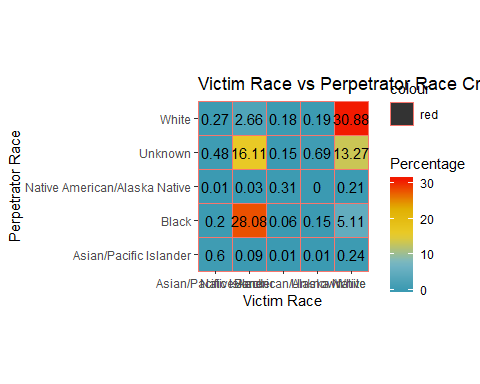


#scale\_y\_discrete(expand = c(0, 0)) +   
 #coord\_equal()

Bivariate Analysis (Categorical+Categorical)

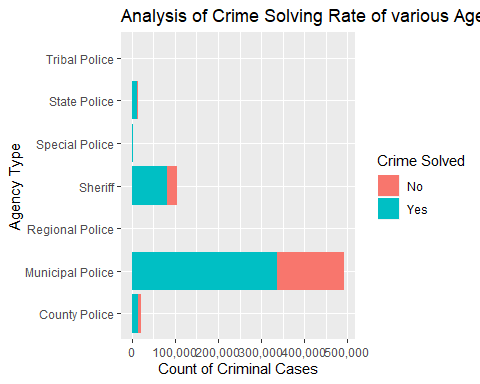
Victim Race vs Perpetrator Race Crime Analysis

homicide\_stats\_race <- homicide\_stats %>%  
 mutate(Total = n()) %>%  
 group\_by(`Victim Race`,`Perpetrator Race`) %>%  
 select(`Victim Race`,`Perpetrator Race`, Total) %>%  
 mutate(Count = n()) %>%  
 mutate(Percentage = round(Count\*100/Total,2)) %>%  
 distinct()  
  
# Gradient color  
grad <- wes\_palette("Zissou1", 100, type = "continuous")  
  
homicide\_stats\_race %>%  
 ggplot(aes(x = `Victim Race`, y = `Perpetrator Race`))+  
 geom\_tile(aes(fill = Percentage, color = "red"))+  
 geom\_text(aes(label = Percentage))+  
 labs(title = "Victim Race vs Perpetrator Race Crime Analysis",  
 x = "Victim Race",  
 y = "Perpetrator Race")+  
 scale\_fill\_gradientn(colours = grad) +   
 scale\_x\_discrete(expand = c(0, 0)) +  
 scale\_y\_discrete(expand = c(0, 0)) +   
 coord\_equal()



Analysis of Crime Solving Rate of various Agencies

homicide\_stats\_t <- homicide\_stats %>%  
 group\_by(`Agency Type`,`Crime Solved`) %>%  
 select(`Agency Type`,`Crime Solved`) %>%  
 mutate(Count = n()) %>%  
 distinct()  
  
  
homicide\_stats\_t %>%  
 ggplot(aes(x = `Agency Type`, y = Count, fill = `Crime Solved`))+  
 geom\_bar(stat = "identity")+  
 coord\_flip()+  
 labs(title = "Analysis of Crime Solving Rate of various Agencies",  
 x = "Agency Type",  
 y = "Count of Criminal Cases")+  
 scale\_y\_continuous(labels = comma)



Conclusion

1. On analysis, we have found out that the crime was prevalent in 1990s. Post that, we have decline in the crime rate.
2. From 1980 to 2014, California is the leading state when it comes to crime, which is followed by Texas.
3. Handgun is the most accessible weapon and thus, is used most in the criminal activities.
4. Below 15 years, most of the criminal cases are reported in case of children between 0-5 years of age for both males and females.
5. First two decades of the dataset, i.e., 1980-1989 and 1990-1999 have the highest crime rates as compared to rest of the years.
6. 49.72% of the criminal cases are reported where the victims belong to white race whereas 46.97% of the cases belong to the victims who are of black race.
7. 30.88% of the cases belong to the category where both the perpetrator and the victim belonged to the white race whereas 28.08% of the cases belong to the category where both are black.
8. Most of the cases are solved by Municipal Police. Most of the unsolved cases are also in the basket of Municipal Police.