Homicides in the U.S.A.

Introduction

This project looks at homicides committed in the United States from 1980 onwards. The analysis is centered around different age-groups and looks at certain sub-populations. It also looks at the evolution of crime levels over time. The random forest model tries to predict the race of perpetrators. The linear model tries to predict the age of the perpetrator based on victim age. The data was downloaded from Kaggle.

About the Dataset

The dataset includes 638k rows and 24 columns. It contains extensive information about homicides such as location, type of crime, murder weapon, relationship between perpetrator and victim as well as age and race information for both victim and perpetrator. The columns which the analysis focuses on are qualitative in nature with the exception of Age.

The dataframe was reduced to 420k rows and 16 columns by removing missing data in certain cells and by removing columns that contained information that was not pertinent to the analysis. The columns Agency_Code, Agency_Name and Agency_Type were dropped as they provided no information for the projections of interest. The rows were reduced because of missing or incomplete data.

Steps Take to Clean Data

Before proceeding with any analysis the Victim_Age, Perpetrator_Age, and Crime_Type columns were cleand. This was done by using the filter() function to retain all the rows within those columns that met the necessary requirements. Thus, Victim ages equal to 998, Perpetrator ages that were blank or less than 4, and Crime type equal to Manslaughter by Negligence were filtered out.

Within the Relationship column, “Boyfriend” and “Girlfriend” were reclassified to “Boyfriend/Girlfriend”, “Employee” and “Employer” to “Colleague”, “Common-Law_Husband” to “Husband”, and “Common-Law_Wife” to “Wife”. This was done by using the $ command to select the Relationship column within the dataset. Then, subsetting was applied by bracket notation to select the categories to change within Relationship. From there, the new categories were applied to the selected categories by vectorization.

Next, the Weapon column was cleaned with the same procedure that was applied to the Relationship column. This allowed to reclassify “Gun”, “Handgun”, “Rifle”, and “Shotgun” to “Firearm”, and “Suffocation” and “Strangulation” to “Suffocation/Strangulation”.

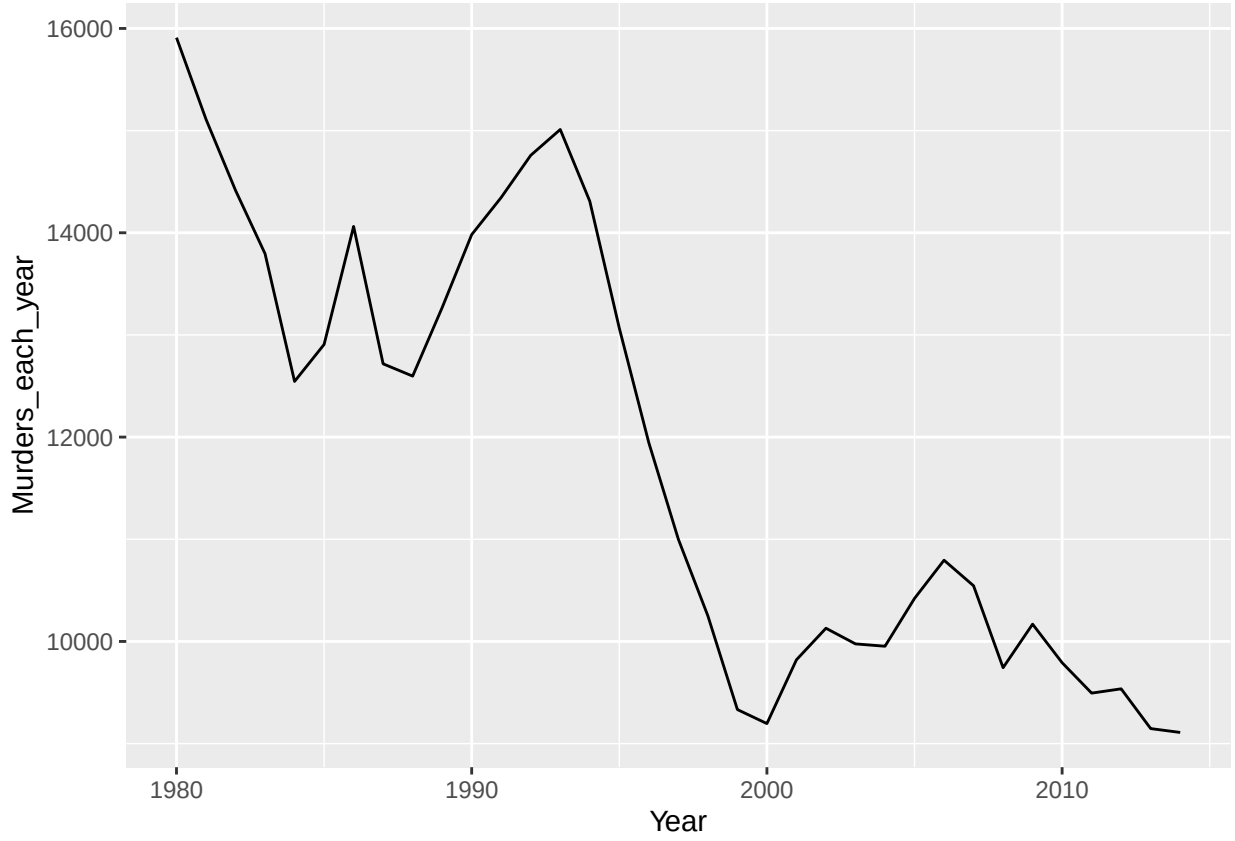
Lastly, new columns of Perpetrator_Age_Group, Victim_Age_Group, and Seasons were created. To create a new column, first a categorize function which includes a single multipath if statement was defined. The if statement makes a logical comparison between the values from the selected column and what is expected by testing for a condition. It only returns a character string when the condition is True. Then, the sapply function was used with the selected column from the dataset and the categorize function as the arguments. The output from sapply was assigned to the new column, which was created by using the $ command on the dataset with the name of the new column. This procedure was used to create the other two new columns.

Time series of murders

Here are the first 5 years of homicides.

Table 1: Evolution of murders over time

Year	Murders_each_year
1980	15910
1981	15111
1982	14415
1983	13795
1984	12545

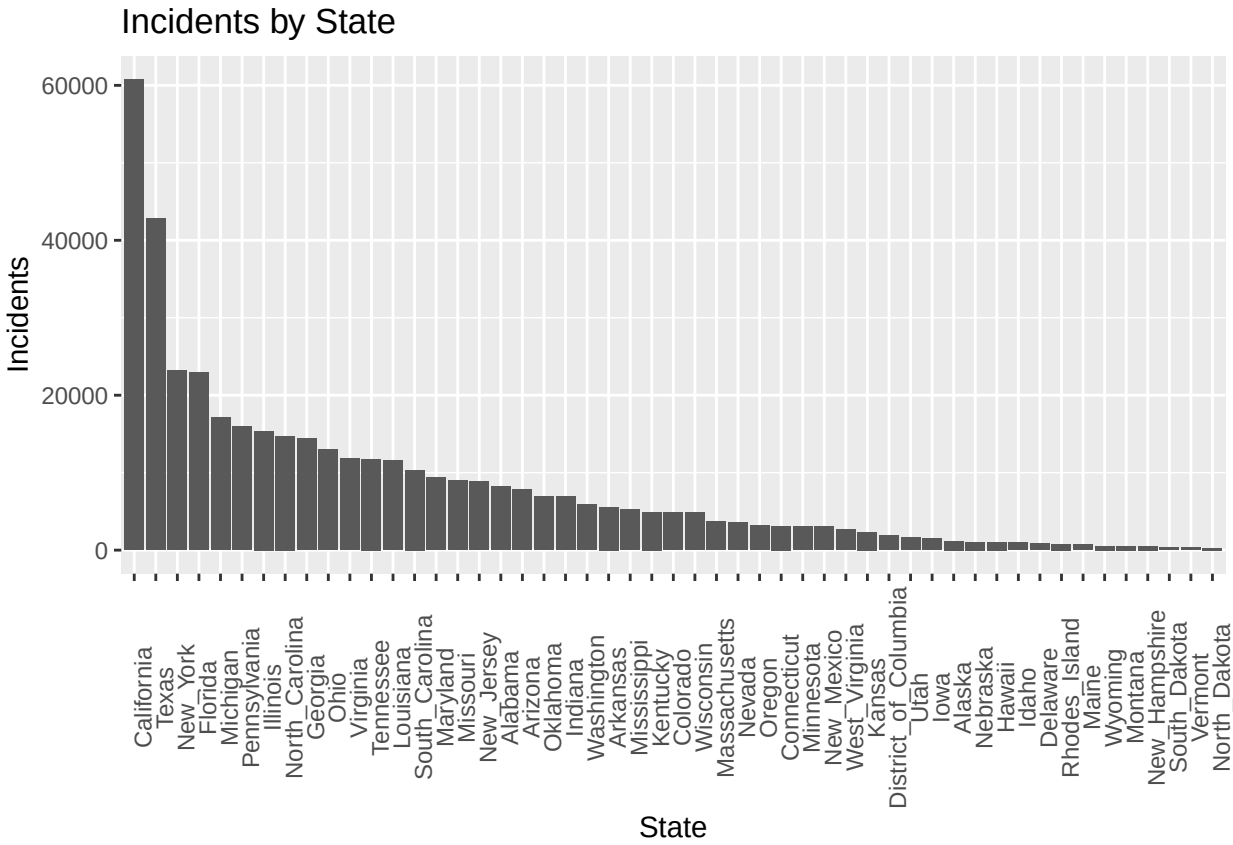


Overview of timeseries

As seen, there is a very sharp drop of murders committed starting in the early 1990's. The worst was at the beginning of the 80's and early 90's

Incidents by State

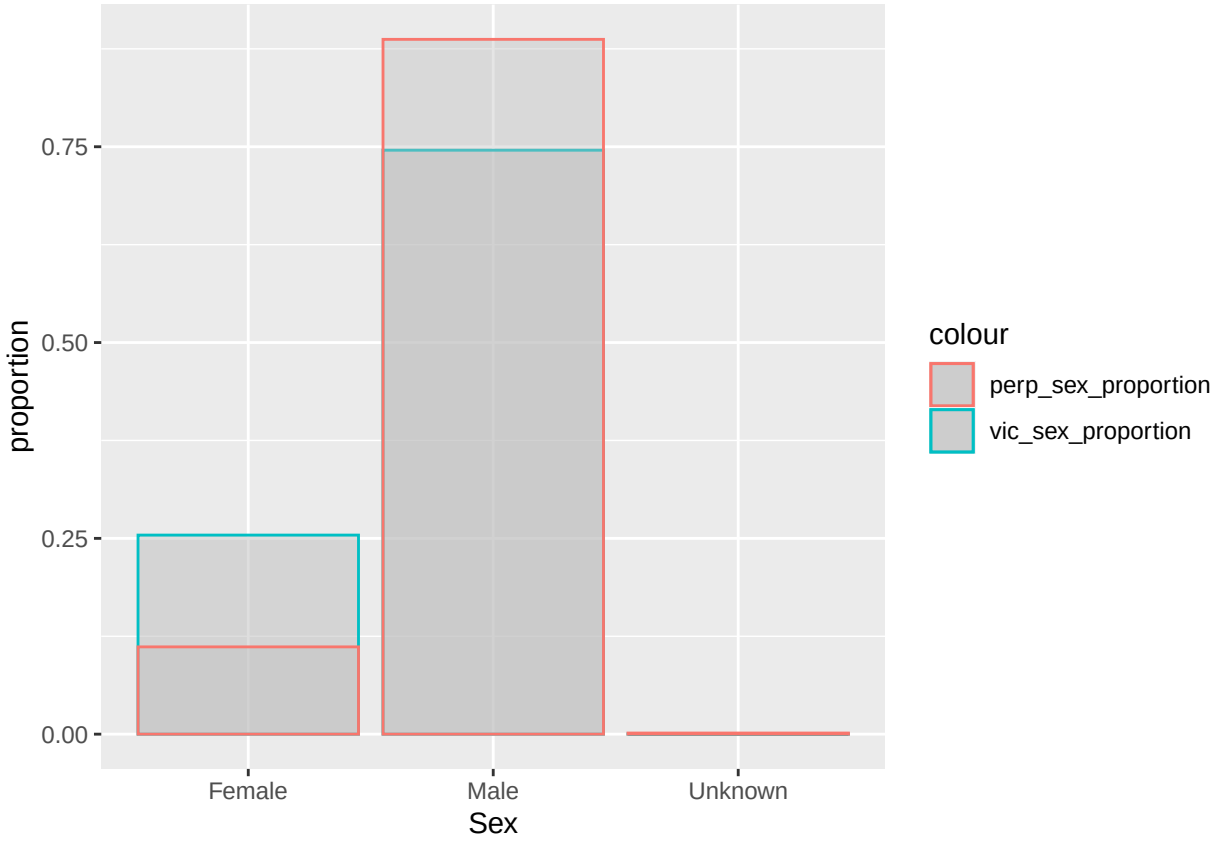
This graph showcases total incidents by state. As expected, the state with the highest population has the most murders. Further into the analysis, data from California was used as it provided a solid foundation for the linear model.



Analysis of Victim vs Perp Sex

Males vs Females

The analysis shows that Females are much more likely to be victims of murders. They are also less likely to be perpetrators.



Analysis of Perpetrators by Age Groups

Adults (18+) vs. Underage (under 18)

Table 2: Adults vs. Underage

Perp_Age_Group	N_homicides	Percent_homicides
Adults	381302	0.922903
Underage	31853	0.077097

Observations of Homicides Committed by Adults vs. Underage

1. Adults perpetrators committed a total of 381302 homicides, which is about 92%.
2. Underage perpetrators committed a total of 31853 homicides, which is about 8%.

Breakdown of Age Groups

Table 3: Breakdown by Perpetrator Age Group

Perpetrator_Age_Group	Count	Proportion
20-29	130738	0.3164381
30-39	91380	0.2211761
40-49	55791	0.1350365
10-19	48607	0.1176483
50-59	30576	0.0740061
9 and under	21702	0.0525275
60-69	16123	0.0390241
70-79	9414	0.0227856
80-89	4515	0.0109281
90 and over	4309	0.0104295

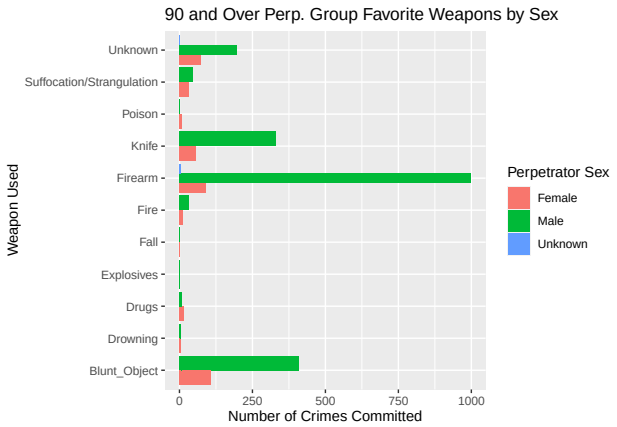
Frequency of Homicides Committed within Age Groups

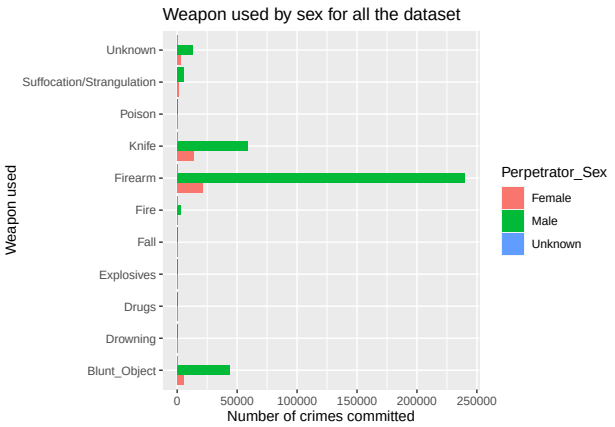
1. The most frequent age group for perpetrators is 20-29, followed by 30-39 and then 10-19.
2. The least frequent age group for perpetrators is 9 and under followed by 90 and over.
3. From the age 40 to 90 the older you are, the least likely you are to commit crime.
4. 264 perpetrators belong in the "90 and over" Perpetrator_Age_Group. It has the second least total perpetrators within all groups.
5. 69972 belong to the group "10-19". It is the 3rd biggest age group in terms of perpetrators. Proportion is 17%.

Elderly (90+) Perpetrators - Weapons of Choice

Table 4: Weapons of Choice by Elderly Perpetrators

Perpetrator_Age_Group	Weapon	Num_Weapons	Proportion
90 and over	Firearm	2043	0.4741239
90 and over	Blunt_Object	891	0.2067765
90 and over	Knife	631	0.1464377
90 and over	Unknown	476	0.1104665
90 and over	Suffocation/Strangulation	124	0.0287770
90 and over	Fire	83	0.0192620
90 and over	Drugs	30	0.0069622
90 and over	Drowning	12	0.0027849
90 and over	Poison	12	0.0027849
90 and over	Explosives	4	0.0009283
90 and over	Fall	3	0.0006962





Observations of Weapons of Choice

1. The most common weapon used was a firearm followed by a blunt object. 179 people used Firearms followed by 33 people using a blunt object.
2. Males favored firearms while females favored blunt objects.

Elderly (90+) Perpetrators - Victims Age

Table 5: Victims Age of Elderly Perpetrators

Perpetrator_Age_Group	Victim_Age_Group	Num_VGroup
90 and over	90 and over	4309

Observations of Victims Age

1. The most targeted Victim_Age_Group was “80-89” with 76 victims within that group.
2. It was followed by “20-29” with 46 victims.

Elderly (90+) Perpetrators - Perpetrators Sex

Table 6: Sex of Elderly Perpetrators

Perpetrator_Age_Group	Perpetrator_Sex	Victim_Age_Group	Weapon	Count
90 and over	Male	90 and over	Firearm	1877
90 and over	Male	90 and over	Blunt_Object	694
90 and over	Male	90 and over	Knife	541
90 and over	Male	90 and over	Unknown	342
90 and over	Female	90 and over	Blunt_Object	197
90 and over	Female	90 and over	Firearm	161
90 and over	Female	90 and over	Unknown	132

Observations of Perpetrators Sex

1. Males killed more people between the ages of “80 - 89” using a firearm (56).

2. Females killed the most people between the ages of “90 and over” using a blunt object(2).
3. Females did not kill anyone in the Victim_Age_Group of “9 and under” as well.
4. Males killed at least 1 person in all Victim_Age_Group categories.

Pre-Teen / Teen Perpetrators (10-19) - Teen Adults

Table 7: Teens Adults

Perpetrator_Age	n
18	18411
19	19846

Observations of Teens Adults

Out of 69972, 38,257 are of age 18 or 19, i.e adults. Therefore 55% of the perpetrators in the group 10-19 are adults.

Pre-Teen / Teen Perpetrators (10-19) - Weapons of Choice

Table 8: Weapons of Choice by Pre-Teen / Teen Perpetrators

Perpetrator_Age_Group	Weapon	Num_Weapons	Proportion
10-19	Firearm	36821	0.7575246
10-19	Knife	6738	0.1386220
10-19	Blunt_Object	2524	0.0519267
10-19	Unknown	1309	0.0269303
10-19	Suffocation/Strangulation	616	0.0126731
10-19	Fire	368	0.0075709
10-19	Drugs	115	0.0023659

Observations of Weapons of Choice

Most common weapon used was a firearm followed by a knife. 48661 used firearm, 10333 knife.

Pre-Teen / Teen Perpetrators (10-19) - Victims Age

Table 9: Victims Age of Pre-Teen / Teen Perpetrators

Perpetrator_Age_Group	Victim_Age_Group	Num_VGroup
10-19	10-19	48607

Observations of Victims Age

1. The most targeted Victim_Age_Group was 20-29.
2. It was followed by 10-19.

Pre-Teen / Teen Perpetrators (10-19) - Perpetrators Sex

Table 10: Sex of Pre-Teen / Teen Perpetrators

Perpetrator_Age_Group	Perpetrator_Sex	Victim_Age_Group	Weapon	Count
10-19	Male	10-19	Firearm	35549
10-19	Male	10-19	Knife	5929
10-19	Male	10-19	Blunt_Object	2416
10-19	Female	10-19	Firearm	1219
10-19	Male	10-19	Unknown	1176
10-19	Female	10-19	Knife	805
10-19	Male	10-19	Suffocation/Strangulation	563
10-19	Male	10-19	Fire	295

Observations of Perpetrators Sex

1. Males killed more people between the ages of "20-29" using a firearm (16652).
2. Females killed more people between the ages of "10-19" using a firearm, followed by the ages "0-9" using a blunt object.

Modeling

Random Forest to Predict Perpetrator's Race

A Random Forest model was created to predict a perpetrator's race based on state, month, perpetrator sex, perpetrator age, weapon, victim sex, victim age, victim race, and relationship. The hypothesis is that the perpetrator's race is very closely related to the victim's race.

As a side note, the ntree = 100 is set for faster run time.

Table 11: Important Variables

Variable	MeanDecreaseAccuracy
Victim_Race	682.13631
Relationship	116.47018
Perpetrator_Age	106.48879
Victim_Age	87.58832
State	80.03135
Weapon	63.09282
Perpetrator_Sex	58.24912
Victim_Sex	40.75394
Month	20.35596

Observations of Random Forest Model - Important Variables

1. Victim_Race is the most important variable, followed by Relationship, Victim_Age, and Perpetrator_Age.
2. The Least important variable is Month.
3. Surprisingly, Victim_Sex is the second least important.

Table 12: Error Details

Perpetrator_Race	Count	Wrong	Error_rate	Overall_error_rate
Asian/Pacific_Islander	1123	487	0.4336598	0.1295519
Black	39176	5418	0.1382990	0.1295519
Native_American/Alaska_Native	708	375	0.5296610	0.1295519
Unknown	692	419	0.6054913	0.1295519
White	40932	4006	0.0978696	0.1295519

Table 13: Confusion Matrix

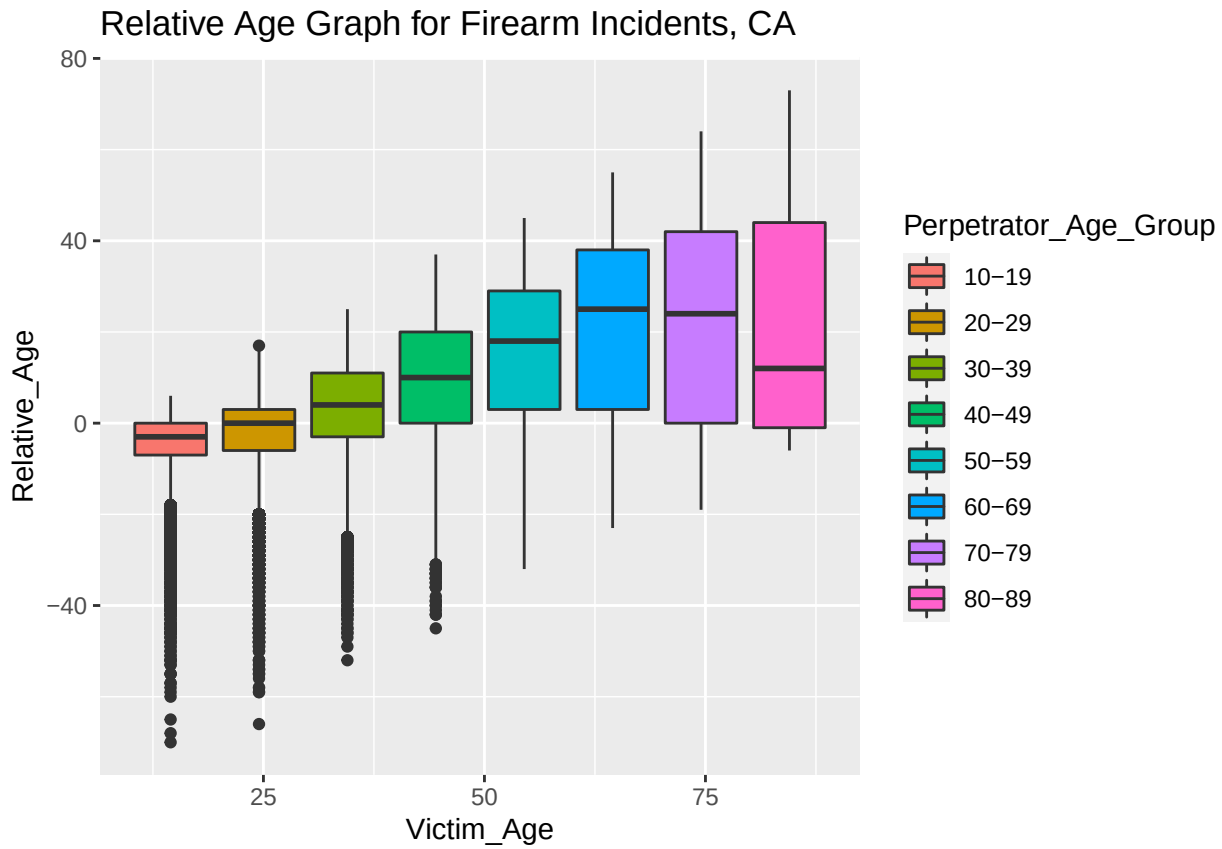
	Asian/Pacific_Islander	Black	Native_American/Alaska_Native	Unknown	Whi
Asian/Pacific_Islander	636	135	3	3	3
Black	97	33752	27	50	52
Native_American/Alaska_Native	6	58	331	1	3
Unknown	8	124	8	273	2
White	194	3622	116	74	369

Observations of Random Forest Model - Error Rate and Confusion Matrix

1. The Error Rate of the model is approximately 0.13.
2. The model is most accurate when Perpetrator_Race = White or Black.
3. The model is least accurate when Perpetrator_Race = Asian/Pacific_Islander, Native_American/Alaska_Native, or Unknown.

Linear Regression to Predict Victim's Age Based on Perpetrator's Age

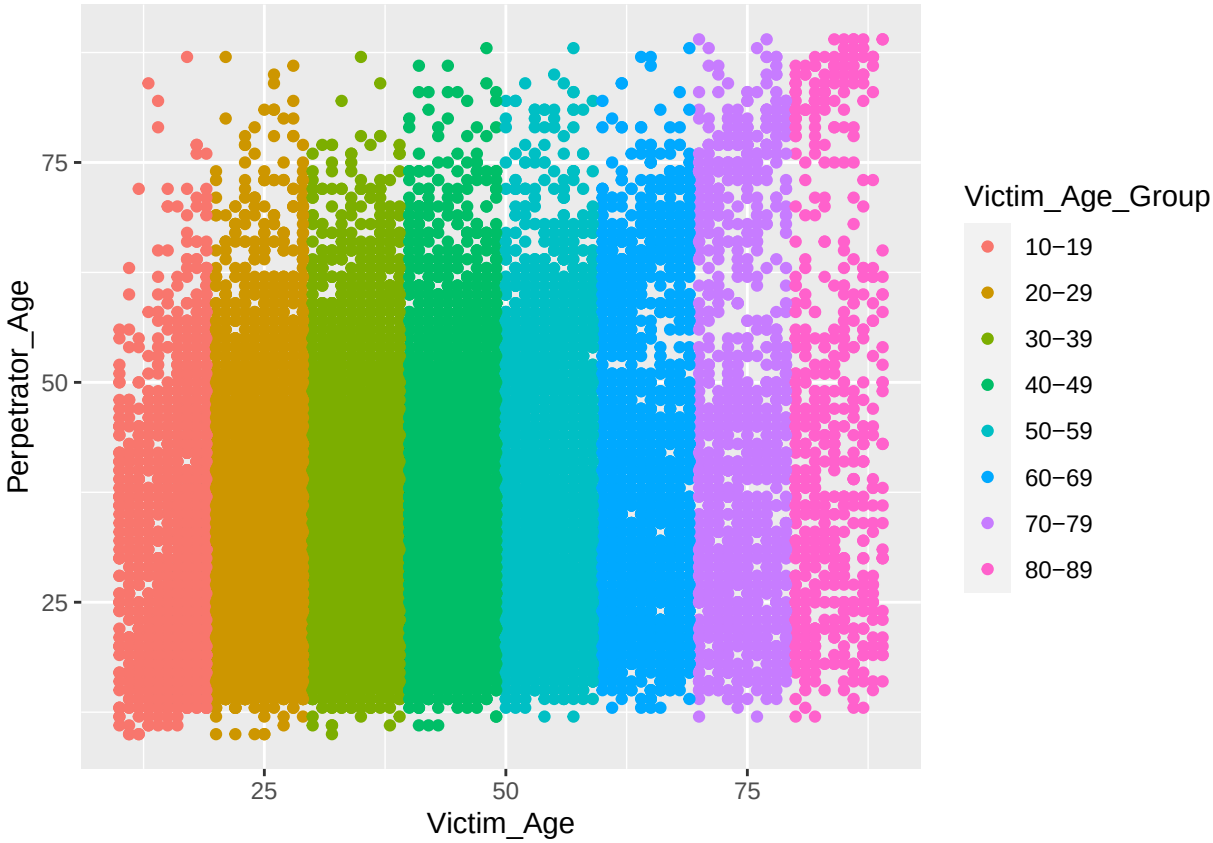
Initially, the relative age (Victim - Perp) is assessed to see if there a pattern. The California data is being used for this analysis.



As observed, up to the age of 40, the victim and perpetrator are very similar in age. The median of each of those groups is almost equal to 0. The older they get the higher the relative age gets; this means that older victims tend to be targeted by individuals younger than they are. One possible interpretation would be that as the pool of potential victims gets older it is difficult to be targeted by someone even older than them, which would lead to a positive relative age. Furthermore, the spread from the boxplot graph gets wider the further we move up the age range. The last age group is mostly outliers as data is sparse and does not follow the same pattern as the rest.

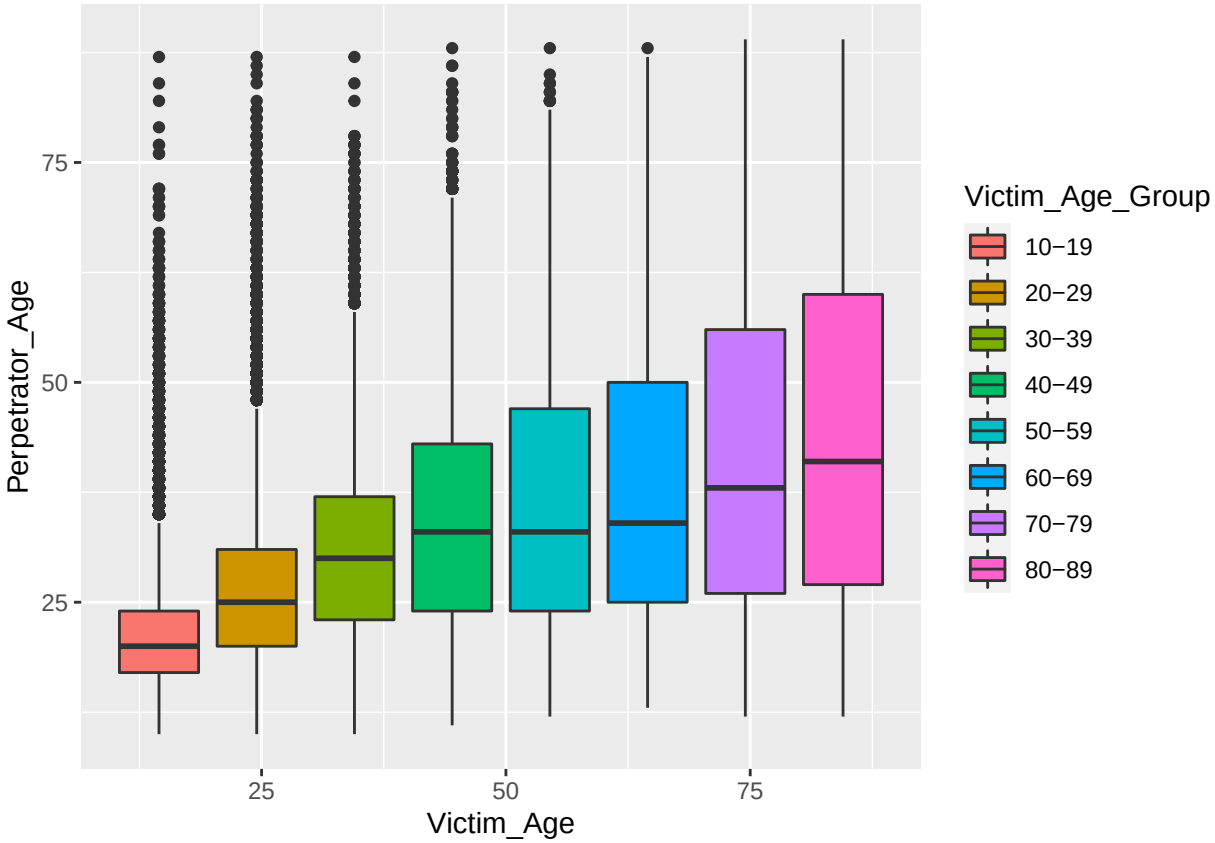
Linear Model

To test the initial hypothesis, that Victim_Age and Perp_Age follow a linear trend, a linear regression model was vreated to predict the perpetrator age based on victim age and other variables for the state of California. Using the data from only one state allowed us to create clearer visuals. The sample used was from California because it had the most murder cases, so there was a lot of data to work with. However, in the plot of Perpetrator_Age by Victim_Age, there are many outliers, specifically for age groups of "9 and under" and "90 and over" between Victim and Perpetrator groups. For ages of 99, this also represents ages greater than 99. So when plotted, a strange pattern for that age is observed. Thus, these age values were filtered out to get a better understanding of the linear relation between Victim_Age and Perpetrator_Age. "9 and under" was also filtered out because there were very few data points in that age group, which skewed the results.



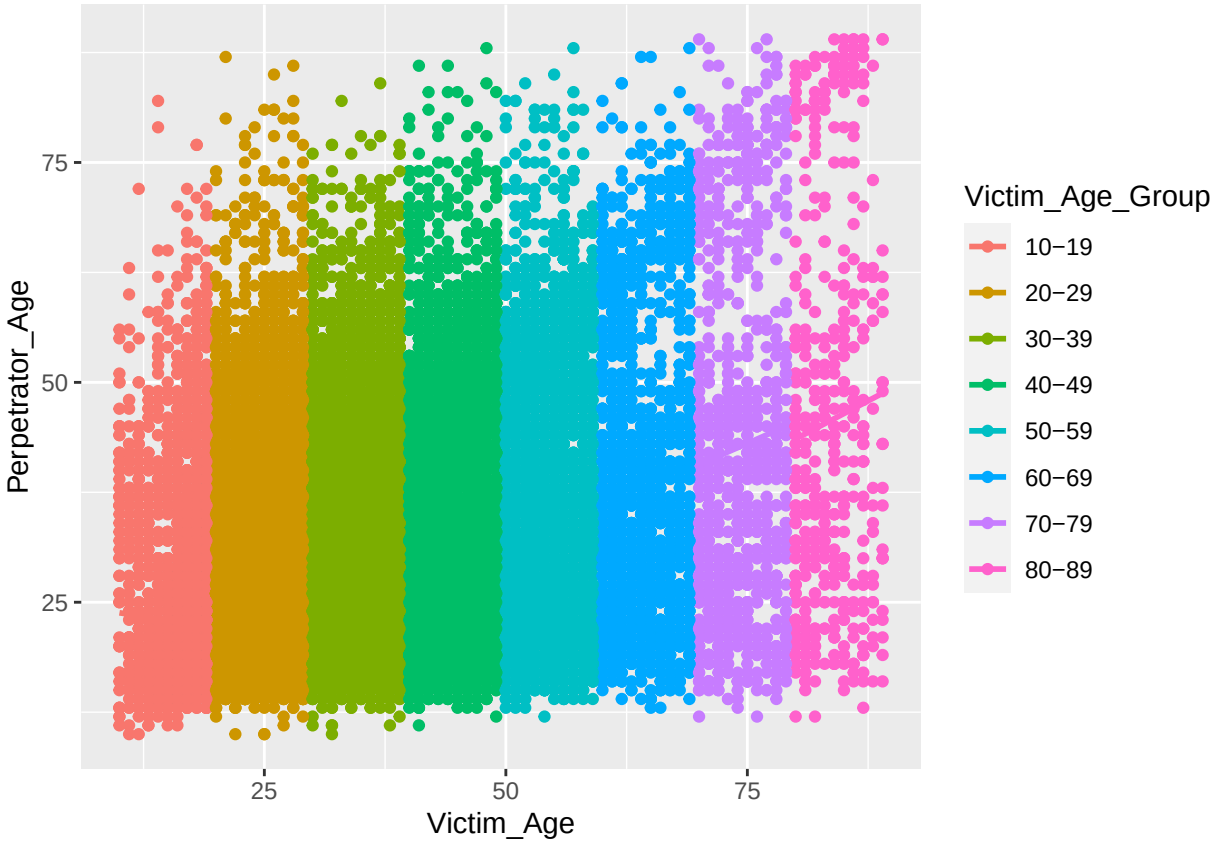
Observations of Age Groups Data - Scatter Plot

1. The Scatter Plot shows the Perpetrator Age by Victim Age relationship for each Victim Age Group.
2. The data looks scattered but you can see a small sign of linear relationship between the two variables. This relationship is not sufficiently clear because there are so many outliers within each Victim groups. To see a clearer linear trend would mean to remove the outliers.



Observations of Age Groups Data - Box Plot

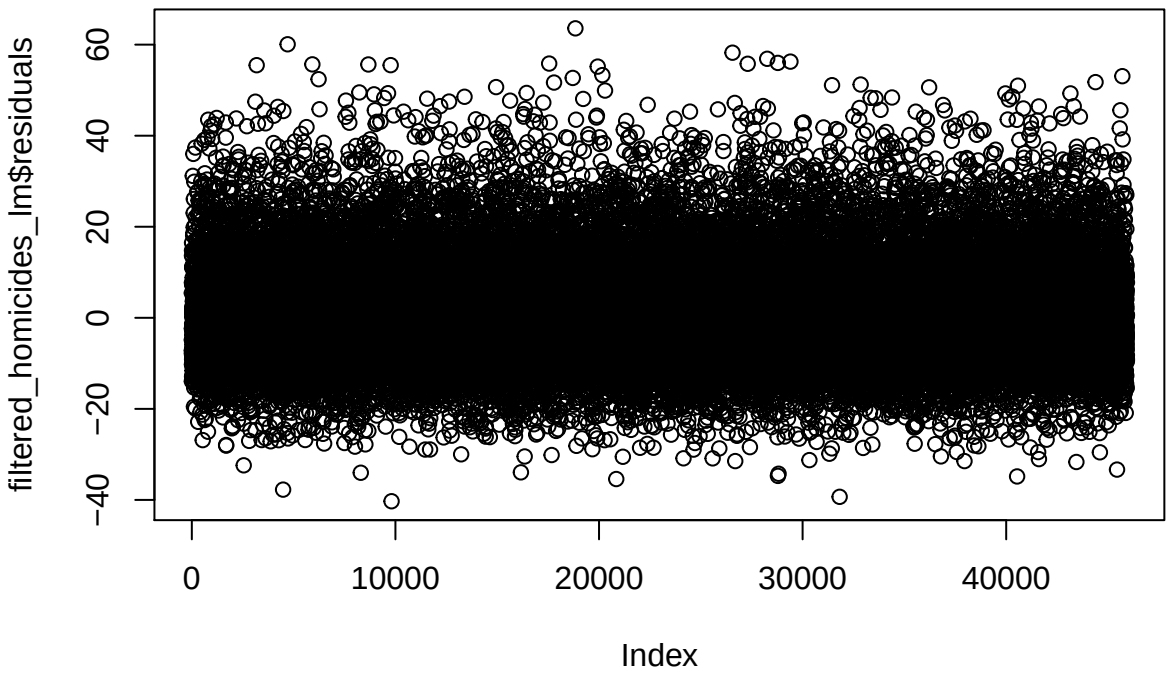
1. As Victim_Age increases, the median Perpetrator_Age increases. This suggest that Victim_Age is a great predictor variable for Perpetrator_Age, since there seems to be a linear regression. It is also observed how the older the Victim is, the wider the spread in Perpetrator_Age is. For example, the median Perpetrator_Age for Victim_Age_Group 20-29 is about 25 years old. This suggests that Perpetrators around 25 years old are more likely to commit a murder within their age range rather than outside it. However, for Victim_Age_Group 80-89, the median Perpetrator_Age is about 40. There is a wider spread of Perpetrator_Age, within this Victim_Age_Group.



Observations of Linear Regression Model

The initial predictor variable was Perpetrator_Age. Then other categorical predictor variables were added based on their p-value to increase the Adjusted R squared. Variables with a p-values greater than 0.05 were removed from the linear model because this meant there was greater probability that the variable will not be meaningful for this regression model. Since they did not improve the Adjusted R-squared, there was no need to keep those variables. In the end, the r-squared is about 0.89.

Since the baseline was removed and the interaction in the linear model included, different slopes for each Victim_Age_Group were obtained. As Perpetrator_Age increases, Victim_Age increases as well. However, many of Victim_Age values outside of the slope of predicted values still exist. This is because Homicide is motivated by a large number of factors making the victim somewhat random, even taking relationship, ages, etc. into account. However, between each Perpetrator_Age_Group there is still a linear pattern between ages of Perpetrators and Victims.



Observations of Linear Regression Model Residuals

A good way to test the accuracy or fit of the model is to look at the residuals or the differences between the real values and the predicted values. The idea here is that the sum of the residuals is approximately zero or close to it as possible. Ideally, when you plot the residuals, they should look random. Otherwise, there might be a hidden pattern that the linear model is not considering.

Looking at the residuals, the majority of it is close to zero. However, some reach 20 and 60 in both directions. Residuals that are 60 could be outliers as there are not many of them. Residuals that fall between 20 and 40 in both directions are concerning, since there are a few. This could indicate that there is a hidden pattern the model cannot capture well. This was expected based on what was previously stated: Victims can be random regardless of their relationship, age, race, etc.

Summary

The analysis conducted, provides a lot of interesting facts and patterns about homicides in the United States. A positive observation is that since 1990 the murder rate has dropped sharply and significantly. As expected, most crimes are committed by adults and the most frequent age group for perpetrators is 20-29 and the most targeted Victim_Age_Group was 20-29. The 20-29 age group represents over a third of the dataset. Moreover, the 90&over do not seem to have a different preference compared to the general population in terms of weapon choice.

Turning to the two models, the random forest offers a low error for the prediction of Perp_Race. The linear model, tries to predict the perpetrator age based on victim age and other variables for the state of California. The regression has a high explanatory power and shows that there is a clear linear trend in the data. The residuals graph also further strengthens the model by not exhibiting any patterns.

Conclusion

Using different analysis methods many different patterns and insights about the homicides in the United States were extracted and observed. The hypothesis was that perpetrators are usually the same race and similar age to that of the victim. Given the analysis conducted, the conclusion is that in an unsolved murder, the perpetrator is likely to be male, of the same race, and similar age group to the victim.