STA 9750 Final Project - US Homicides 1980-2014

Amy Xu, David Genfan, Davin Yu and Farwa Ismail

5/24/2020

Data Summary

This dataset was pulled from kaggle. Founded by Thomas Hargrove, the Murder Accountability Project is the most complete database of homocides in the United States. It spans from 1980 to 2014 and includes variables such as the age, race, sex, ethnicity of the victims and perpetrators, as well as their relationships and the weapons used. The data was sourced from the FBI's Supplementary Homocide Report and Freedom of Information Act (FOIA) requests.

Below are all the variables we have in our dataset:

```
##  [1] "record.id"             "agency.code"           "agency.name"
##  [4] "agency.type"           "city"                  "state"
##  [7] "year"                  "month"                 "incident"
## [10] "crime.type"            "crime.solved"          "victim.sex"
## [13] "victim.age"            "victim.race"           "victim.ethnicity"
## [16] "perpetrator.sex"       "perpetrator.age"       "perpetrator.race"
## [19] "perpetrator.ethnicity" "relationship"          "weapon"
## [22] "victim.count"          "perpetrator.count"     "record.source"
```

The summary for the dataset is included in the appendix which shows what each column looks like. In addition, shown below are some summary points:

Attributes	Values
Dimensions	638454, 24
No. of Unique Record IDs	0
Any Record IDs Duplicated?	0
Any NA values in the database?	TRUE

Cleaning the Dataset

Looking further into the dataset, we see that `victim.age` has a value of 998 for some records. It is fair to say that humans don't live that long, therefore we will filter the dataset to not include these records. Moreover, we ignore `perpetrator.age` below 18 since laws are different for juviniles and some ages go as low as zero which doesn't make sense.

Then there is the issue of 'Unknown' sexes, races and ethnicities. We noticed that `victim.ethnicity` and `perpetrator.ethnicity`, both have a lot of Unknown values. Along with that, these variables only include whether the individual was hispanic or not. So for the most part, we will be ignoring these variables.

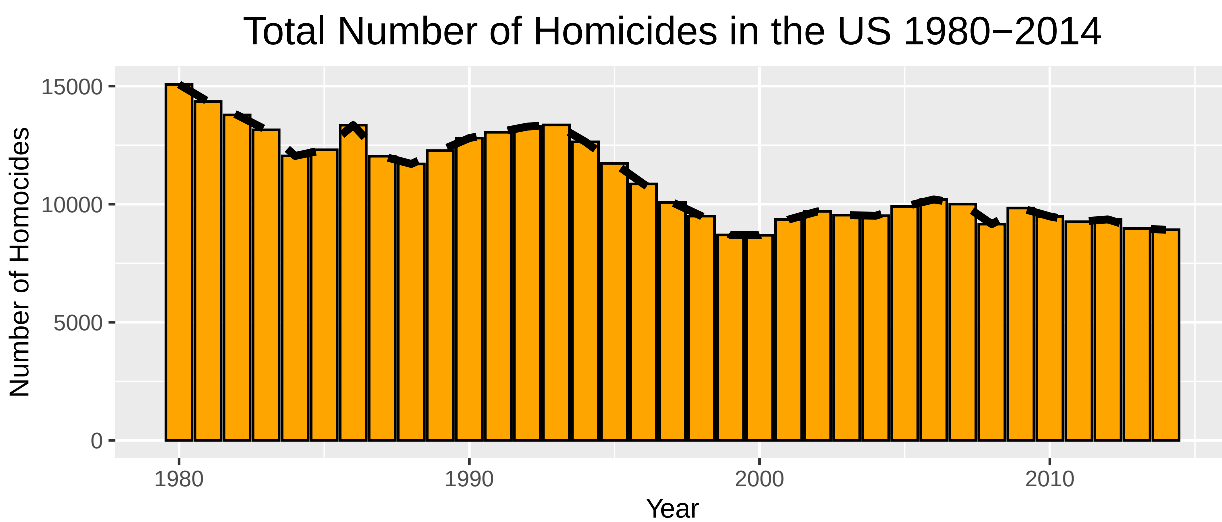
`incident`, `victim.count` and `perpetrator.count` are three more variables which don't make sense. Not much has been said about them by the sources of the dataset as well.

After filtering for `victim.age`, and removing missing records, we have 387844 records left.

Exploratory Analysis

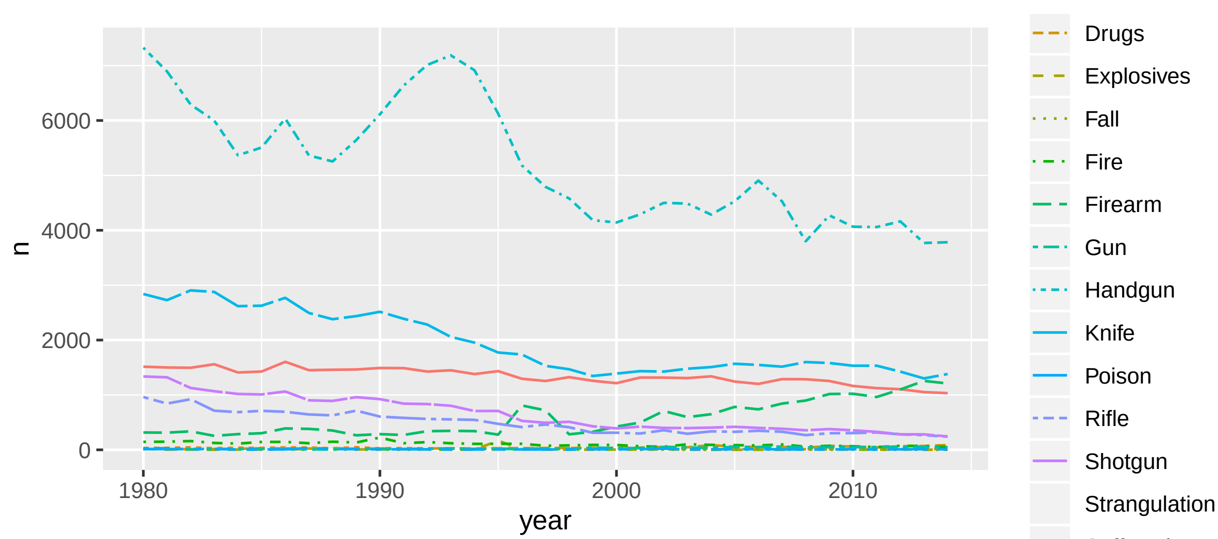
1. Homicide over the years 1980-2014

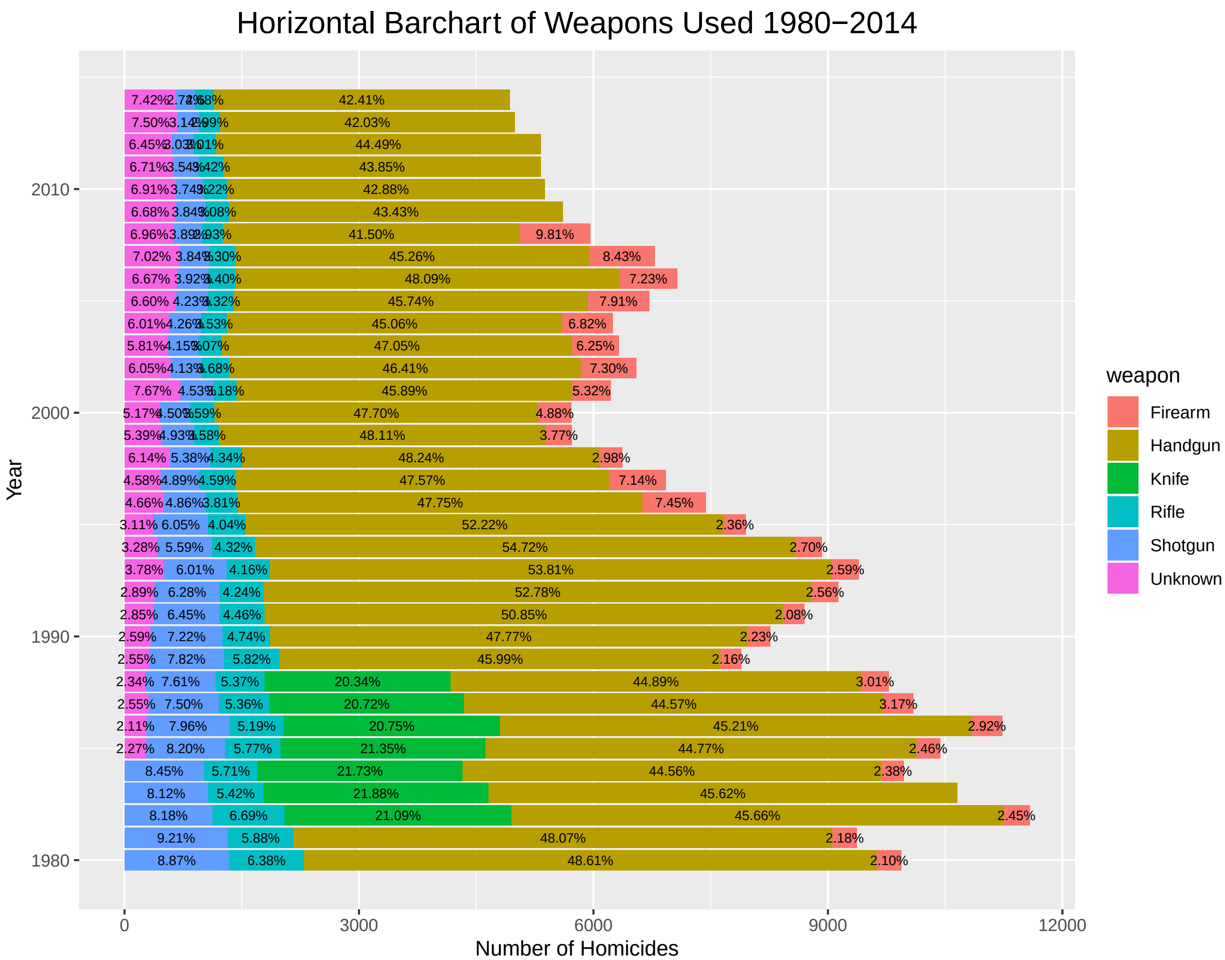
As a starting point, we looked at the total number of homicides per year from 1980 to 2014. The number of homicides peaked in 1993 and declined sharply until 1999. From then, it rose gradually until 2007 and declined thereafter. Overall, homicides are down from the levels experienced in the 1990s. Why did the number of homicides decline so sharply from 1993 to 1999?



2. Weapon use over the years 1980-2014

To explore that question, we plotted the weapons used from 1980-2014. It's a little difficult to distinguish the different weapons given the sixteen classifications involved.

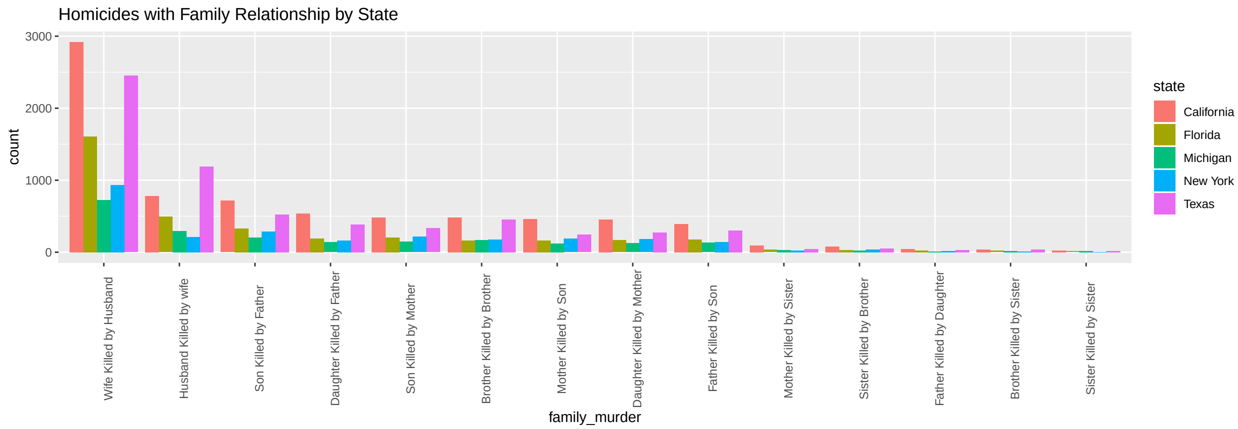




By including only those weapons that have appeared 2% of the time or more, we narrowed that list to: Blunt Object, Firearm, Handgun, Knife, Rifle, Shotgun, Unknown. The horizontal bar chart plots the make-up of homicides each year by weapon-type. In 1993, 57.69% of homicides were committed with a handgun. If we include other gun-related weapon types such as firearm, rifle or shotgun, that figure rises. But, overall, hand-gun related homicides are down from the levels experienced in the 1990s.

3. Perpetrator and victim relationship status (Family) by State.

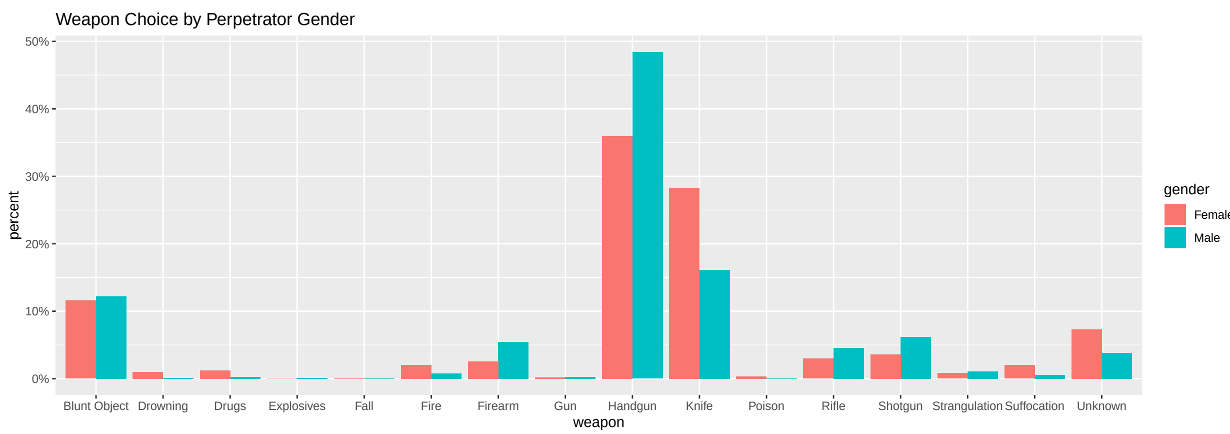
We are also interested in the relationships between victims and perpetrators, specifically those where the relationship is familial. First we filter to keep only the cases where the relationship status indicates an immediate relationship: Father, Mother, Brother, Sister, Son, Daughter, Wife, Husband. Analysis will focus on the 3 states with the most crimes solved: California, Texas, and New York.



In all three states, homicides involving a family relationship show that the wife being killed by the husband tower is the most common. In Texas and California, husband killed by wife is the second most common relationship. However, in New York wife killed by husband is followed by the son killed by father and son killed by mother. Sisters killed by sisters are the least prevalent in all three states.

4. Percentage of weapons used by perpetrator gender.

We also examined weapon choice by perpetrator gender. We found that females used blunt objects and knives more then males. The weapon of choice for males is dominated by Handgun.



Modeling

Some Initial Cleaning for Modeling

If we look at the types of variables we have available to us, we find that a lot of them are categorical, and often in this case the more categories and possible combinations we have, the more sparse our data will become for any sort of inference. We would like to maximize the data available for any particular combination of categorical variables, so as to make the inference process more accurate. In this vein, we will combine some of our categories to reduce the total number of possible combinations.

With some foresight as to the models we are about to run, weapons used in a homicide will be reduced to larger brackets such as: "guns","household objects", "physical force", "chemicals" and an "unknown" category. In addition, relationships are reduced to those part of the victim's family and those that are not.

We will also perform a 70 train-test split, and get rid of categorical variables with large number of factors like ‘cities’, which has 1768 unique cities.

Who’s the Killer?

In this section we aim to make a classifier that can predict whether the perpetrator was a part of the victim’s family, based on some information about the homicide case. This, if accurate to some extent, can be useful in helping law enforcement see what the data says about the perpetrator most standard cases, since this uses correlations based on actual past data.

Our problem is then a classification problem of predicting a logical variable $Q_{Familiy}$ defined as:

$$Q_{Family} = \begin{cases} 1 & \text{if perpetrator was part of the victim's family} \\ 0 & \text{if perpetrator was not in the family} \end{cases}$$

against multiple variables that we might have. As of this point in the analysis we have kept the following columns for our analysis. It’s natural at this point to ask: **what are our features?**, what will we use to predict Q_{Family}?

```
##  [1] "agency.type"          "crime.type"           "crime.solved"
##  [4] "victim.sex"           "victim.age"           "victim.race"
##  [7] "perpetrator.sex"      "perpetrator.age"      "perpetrator.race"
## [10] "record.source"        "weapon_grouped"       "relationship_grouped"
```

Our approach to constructing this classifier is to try a few different models that lend themselves well to the structure of this problem, like `glm`, `naive bayes` and finally a `random forest classifier` and see which models perform best. The process that follows was a mix of running different models and finding which features were *useful* along the way.

Generalised Linear Model

The first thing to try naturally was a `glm`. We have a classification problem and we felt it would be great to see what kind of relations there are in the data in one cheap and easy go. Initially we would have liked to use the `stepwise` function to help us in feature selection - getting the best fit while minimizing overfitting. However, it turned out that using the stepwise function to run many logistic regressions (as was the case with our particular classification problem) caused issues and the stepwise function would simply not converge. We chose not to follow that direction and instead went with another approach.

We started by running a simple logisitic regression on variables that made sense to us, to first see what kind of results we get. It would make sense to look at the sex and age of the victim and perpetrator as well as what kind of weapon was used. Our logistic regression problem looked like:

$$Q_{family} \sim \text{victim.sex} + \text{perpetrator.sex} + \text{victim.age} + \text{perpetrator.age} + \text{weapon}$$

We found an accuracy of the above generalised linear model to be 80.4070034 %. Which is decent. The confusion matrix is also shown below:

	FALSE	TRUE
FALSE	58732	9642
TRUE	10904	25586

There are a sizable amount of false positives and false negatives, but we would like to see if there's any improvement if we bring in the rest of our variables, and is the increase in parameters *worth it?*

So we run a new model, and this time with all of the variables we made available, not just the ones that seem natural, and we compare the `AIC` of these models.

	df	AIC
model1	12	199989.4
model2	29	195661.1

The accuracy of that model was 80.6311031 % and the confusion matrix is shown below.

	FALSE	TRUE
FALSE	58732	9642
TRUE	10904	25586

We find that the `AIC` value of our initial glm is 1.9998936×10^5, which is greater than the `AIC` value of our second model with a value of 1.9566115×10^5. This means that having all those extra factors is worth the increase in parameters and we can feel relatively safe about not overfitting with our number of parameters.

Naive Bayes

Another model in our arsenal is a simple yet powerful statistical tool for looking at the correlations hidden in categorical data: The Naive Bayes Model. The fact that we have a large dataset with a good amount of categorical variables makes this model well suited to be used in our analysis. We run the analysis with some laplace smoothing to make sure that any combination of categorical variables that are spread too thin are analysed appropriately. The summary for this model is shown below.

```
##
## ================================== Naive Bayes ==================================
##
## - Call: naive_bayes.formula(formula = relationship_grouped ~ ., data = train,      laplace = 3)
## - Laplace: 3
## - Classes: 2
## - Samples: 209728
## - Features: 11
## - Conditional distributions:
##     - Bernoulli: 3
##     - Categorical: 6
##     - Gaussian: 2
## - Prior probabilities:
##     - FALSE: 0.6624
##     - TRUE: 0.3376
##
## ---------------------------------------------------------------------------------
```

More importantly, we find that the Naive Bayes algorithm is worse than our glm models. It gives us an accuracy of 75.0972688%, and has the following confusion matrix.

	FALSE	TRUE
FALSE	59079	15557
TRUE	10557	19671

Even though the interpretability of such a model makes it very attractive to consider, we feel that it is not worth the decrease in accuracy when compared to the `glm`.

Random Forest Model

Our last attempt will be on a random forest model. These are quite general and flexible models and are hence quite attractive to use in this situation, though they may take a longer time to run.

```
## ntree      OOB      1      2
##    10:  17.31% 12.30% 27.14%
##    20:  16.83% 12.22% 25.86%
##    30:  16.72% 12.21% 25.55%
##    40:  16.67% 12.21% 25.43%
##    50:  16.63% 12.26% 25.20%
##    60:  16.58% 12.26% 25.07%
##    70:  16.59% 12.27% 25.05%
##    80:  16.58% 12.29% 25.00%
##    90:  16.58% 12.29% 25.02%
##   100:  16.56% 12.29% 24.92%
##   110:  16.57% 12.30% 24.94%
##   120:  16.55% 12.31% 24.87%
##   130:  16.54% 12.31% 24.83%
##   140:  16.53% 12.30% 24.82%
##   150:  16.54% 12.34% 24.79%
##   160:  16.53% 12.36% 24.70%
##   170:  16.54% 12.38% 24.71%
##   180:  16.53% 12.38% 24.67%
##   190:  16.54% 12.36% 24.73%
##   200:  16.53% 12.36% 24.72%
##   210:  16.52% 12.35% 24.70%
##   220:  16.52% 12.36% 24.69%
##   230:  16.53% 12.37% 24.68%
##   240:  16.53% 12.36% 24.71%
##   250:  16.52% 12.35% 24.69%
##   260:  16.53% 12.37% 24.69%
##   270:  16.53% 12.36% 24.72%
##   280:  16.53% 12.36% 24.71%
##   290:  16.52% 12.37% 24.67%
##   300:  16.52% 12.37% 24.68%
##   310:  16.52% 12.37% 24.66%
##   320:  16.52% 12.37% 24.64%
##   330:  16.51% 12.38% 24.62%
##   340:  16.52% 12.38% 24.62%
##   350:  16.52% 12.38% 24.64%
##   360:  16.53% 12.39% 24.64%
##   370:  16.52% 12.38% 24.63%
##   380:  16.52% 12.39% 24.60%
##   390:  16.52% 12.39% 24.62%
```

```
##   400:  16.52% 12.40% 24.61%
##   410:  16.53% 12.40% 24.62%
##   420:  16.52% 12.41% 24.59%
##   430:  16.52% 12.40% 24.60%
##   440:  16.52% 12.41% 24.59%
##   450:  16.52% 12.42% 24.58%
##   460:  16.53% 12.41% 24.60%
##   470:  16.52% 12.40% 24.59%
##   480:  16.52% 12.41% 24.59%
##   490:  16.52% 12.42% 24.56%
##   500:  16.52% 12.42% 24.57%
```

```
##                 Length Class  Mode
## call                 6 -none- call
## type                 1 -none- character
## predicted       209728 factor numeric
## err.rate          1500 -none- numeric
## confusion            6 -none- numeric
## votes           419456 matrix numeric
## oob.times       209728 -none- numeric
## classes              2 -none- character
## importance          44 -none- numeric
## importanceSD        33 -none- numeric
## localImportance      0 -none- NULL
## proximity            0 -none- NULL
## ntree                1 -none- numeric
## mtry                 1 -none- numeric
## forest              14 -none- list
## y               209728 factor numeric
## test                 0 -none- NULL
## inbag                0 -none- NULL
## terms                3 terms  call
```

We get the following confusion matrix for the Random Forest Model:

	FALSE	TRUE
FALSE	61171	8677
TRUE	8465	26551

The accuracy of 83.6531126% and comparatively smaller off diagonal elements on the confusion matrix make the random forest model the most accurate so far, and well worth the increase in computation time.

Model Validation

We compile the results and compare our different models below.

Model	Accuracy
Model 1 (GLM)	0.8040700
Model 2 (GLM)	0.8063110
Model nb (Naive Bayes)	0.7509727
Model rf (Random Forest)	0.8365693

Based on the accuracy we are getting on our `test` set, we see that the most useful algorithm so far is the **random forest algorithm**.

Final Words It remains to be seen if any results from such a model are actually useful in the real world. Many inherent biases may show up in the data, and results from such data are only as biased as the data collection method. Through out our analysis the source of the data may be transparent, but the collection process is opaque and heterogenously managed as most datasets are in the real world. The sensitive nature of the topic at hand lends us to be careful of any implications extracted from such an analysis without looking into further detail, the collection methods and the anthropological mechanisms surrounding them.

Appendix

Summary of the dataset

```
##    record.id       agency.code          agency.name
##  Min.   :     2   NY03030: 14337   Los Angeles : 15976
##  1st Qu.:145927   CA01942: 13832   New York    : 14337
##  Median :311879   ILCPD00: 10773   Chicago     : 10773
##  Mean   :313864   MI82349:  8416   Detroit     :  8416
##  3rd Qu.:479174   TXHPD00:  8121   Houston     :  8263
##  Max.   :638454   PAPEP00:  7531   Philadelphia:  7544
##                   (Other):324834   (Other)     :322535
##            agency.type               city                state
##  County Police    : 13229   Los Angeles : 23157   California  : 56181
##  Municipal Police:286425   New York    : 14346   Texas       : 40198
##  Regional Police :   174   Cook        : 11438   Florida     : 22821
##  Sheriff          : 75275   Harris      : 10308   New York    : 20973
##  Special Police   :  1838   Wayne       : 10132   Michigan    : 15680
##  State Police     : 10860   Philadelphia:  7532   Pennsylvania: 15191
##  Tribal Police    :    43   (Other)     :310931   (Other)     :216800
##       year            month           incident
##  Min.   :1980   July     : 35608   Min.   :  0.00
##  1st Qu.:1987   August   : 35068   1st Qu.:  1.00
##  Median :1994   May      : 32917   Median :  2.00
##  Mean   :1996   January  : 32895   Mean   : 20.75
##  3rd Qu.:2004   June     : 32706   3rd Qu.:  6.00
##  Max.   :2014   September: 32676   Max.   :999.00
##                 (Other)  :185974
##                        crime.type     crime.solved   victim.sex
##  Manslaughter by Negligence:  6542   No :   574   Female :101368
##  Murder or Manslaughter    :381302   Yes:387270   Male   :286368
##                                                   Unknown:   108
##
##
##
##
##    victim.age                            victim.race        victim.ethnicity
##  Min.   : 0.00   Asian/Pacific Islander       :  6030   Hispanic    : 40004
##  1st Qu.:22.00   Black                        :167916   Not Hispanic:129286
##  Median :30.00   Native American/Alaska Native:  3240   Unknown     :218554
##  Mean   :33.77   Unknown                      :  2905
##  3rd Qu.:42.00   White                        :207753
```

```
##  Max.   :99.00
##
##  perpetrator.sex  perpetrator.age                      perpetrator.race
##  Female : 44865   Min.   :18.00   Asian/Pacific Islander       :  5345
##  Male   :342437   1st Qu.:22.00   Black                        :179800
##  Unknown:   542   Median :29.00   Native American/Alaska Native:  3223
##                   Mean   :32.05   Unknown                      :  3237
##                   3rd Qu.:38.00   White                        :196239
##                   Max.   :99.00
##
##   perpetrator.ethnicity       relationship             weapon
##  Hispanic    : 39811    Acquaintance:110663   Handgun     :181970
##  Not Hispanic:129470    Unknown     : 73252   Knife       : 67937
##  Unknown     :218563    Stranger    : 72479   Blunt Object: 46927
##                         Wife        : 22840   Shotgun     : 22919
##                         Friend      : 18942   Firearm     : 19900
##                         Girlfriend  : 15874   Rifle       : 17025
##                         (Other)     : 73794   (Other)     : 31166
##   victim.count     perpetrator.count record.source
##  Min.   : 0.0000   Min.   : 0.0000   FBI :374778
##  1st Qu.: 0.0000   1st Qu.: 0.0000   FOIA: 13066
##  Median : 0.0000   Median : 0.0000
##  Mean   : 0.1363   Mean   : 0.2204
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :10.0000   Max.   :10.0000
##
```

```
##  [1] "record.id"             "agency.code"           "agency.name"
##  [4] "agency.type"           "city"                  "state"
##  [7] "year"                  "month"                 "incident"
## [10] "crime.type"            "crime.solved"          "victim.sex"
## [13] "victim.age"            "victim.race"           "victim.ethnicity"
## [16] "perpetrator.sex"       "perpetrator.age"       "perpetrator.race"
## [19] "perpetrator.ethnicity" "relationship"          "weapon"
## [22] "victim.count"          "perpetrator.count"     "record.source"
```

```
##             record.id           agency.code           agency.name
##             "integer"              "factor"              "factor"
##           agency.type                  city                 state
##              "factor"              "factor"              "factor"
##                  year                 month              incident
##             "integer"              "factor"             "integer"
##            crime.type          crime.solved            victim.sex
##              "factor"              "factor"              "factor"
##            victim.age           victim.race      victim.ethnicity
##             "integer"              "factor"              "factor"
##       perpetrator.sex       perpetrator.age      perpetrator.race
##              "factor"             "integer"              "factor"
## perpetrator.ethnicity          relationship                weapon
##              "factor"              "factor"              "factor"
##          victim.count     perpetrator.count         record.source
##             "integer"             "integer"              "factor"
```

```
## Observations: 387,844
```

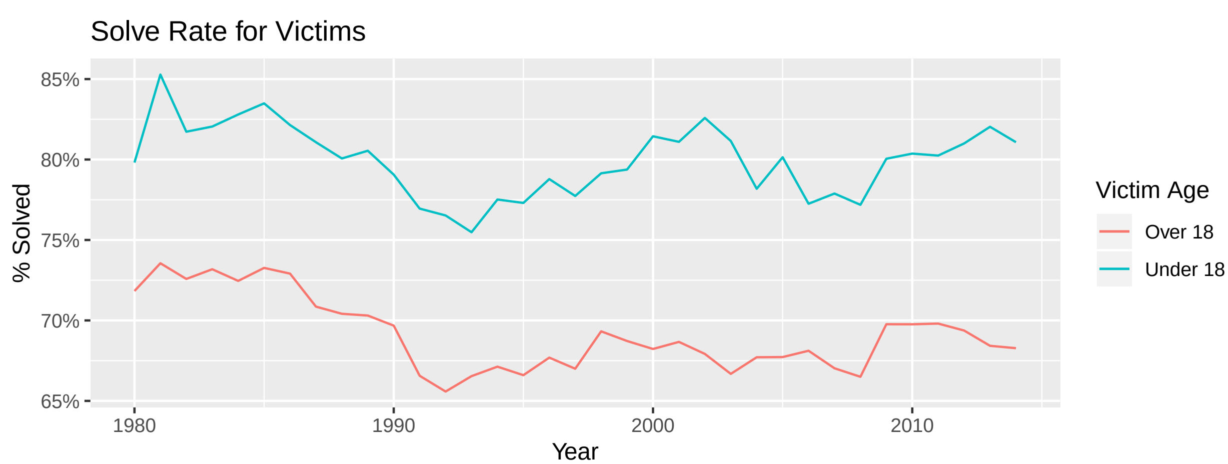
```
## Variables: 24
## $ record.id             <int> 2, 4, 6, 7, 8, 10, 12, 13, 14, 16, 17, 18, 19...
## $ agency.code           <fct> AK00101, AK00101, AK00101, AK00101, AK00101, ...
## $ agency.name           <fct> Anchorage, Anchorage, Anchorage, Anchorage, A...
## $ agency.type           <fct> Municipal Police, Municipal Police, Municipal...
## $ city                  <fct> Anchorage, Anchorage, Anchorage, Anchorage, A...
## $ state                 <fct> Alaska, Alaska, Alaska, Alaska, Alaska, Alask...
## $ year                  <int> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 198...
## $ month                 <fct> March, April, May, May, June, June, July, Jul...
## $ incident              <int> 1, 1, 1, 2, 1, 3, 2, 3, 1, 3, 1, 1, 1, 1, 1, ...
## $ crime.type            <fct> Murder or Manslaughter, Murder or Manslaughte...
## $ crime.solved          <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, ...
## $ victim.sex            <fct> Male, Male, Male, Female, Female, Male, Male,...
## $ victim.age            <int> 43, 43, 30, 42, 99, 38, 20, 36, 20, 31, 16, 3...
## $ victim.race           <fct> White, White, White, Native American/Alaska N...
## $ victim.ethnicity      <fct> Unknown, Unknown, Unknown, Unknown, Unknown, ...
## $ perpetrator.sex       <fct> Male, Male, Male, Male, Male, Male, Male, Mal...
## $ perpetrator.age       <int> 42, 42, 36, 27, 35, 40, 49, 39, 49, 29, 19, 2...
## $ perpetrator.race      <fct> White, White, White, Black, White, Unknown, W...
## $ perpetrator.ethnicity <fct> Unknown, Unknown, Unknown, Unknown, Unknown, ...
## $ relationship          <fct> Acquaintance, Acquaintance, Acquaintance, Wif...
## $ weapon                <fct> Strangulation, Strangulation, Rifle, Knife, K...
## $ victim.count          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, ...
## $ perpetrator.count     <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...
## $ record.source         <fct> FBI, FBI, FBI, FBI, FBI, FBI, FBI, FBI, FBI, ...
```

Other code used to explore dataset

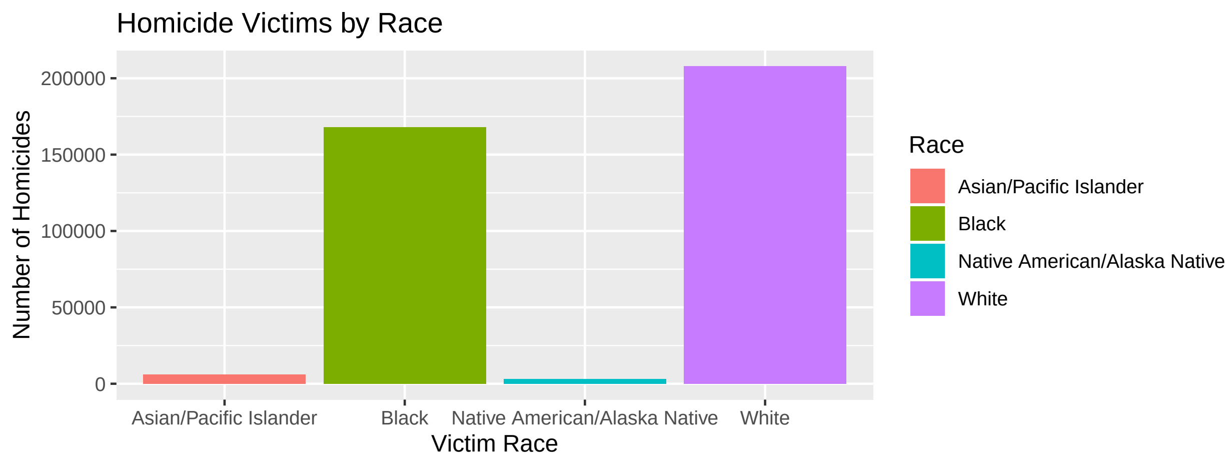
```
## [1] Unknown      Not Hispanic Hispanic
## Levels: Hispanic Not Hispanic Unknown
```

```
## [1] Unknown      Not Hispanic Hispanic
## Levels: Hispanic Not Hispanic Unknown
```

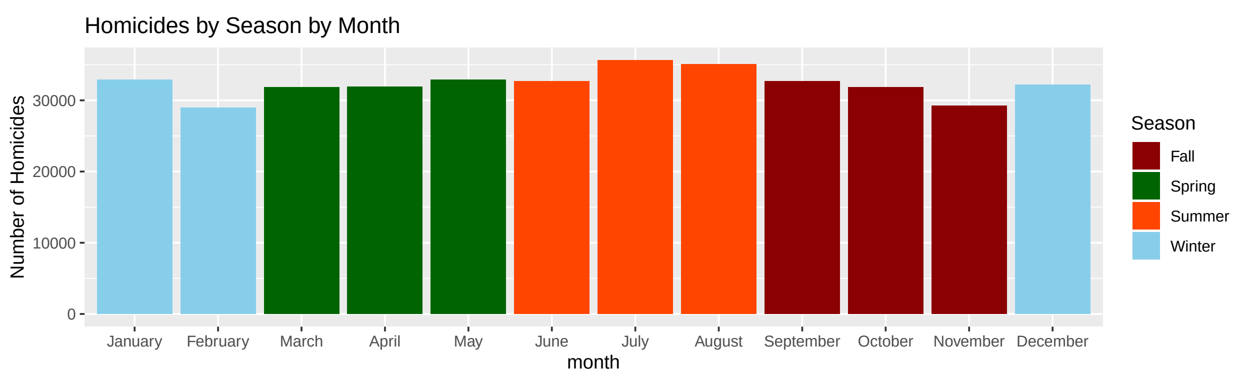
Solve Rate based on victims age.



Homicide Victims based on Race



Homicide count based on Season and Month



Summer months: July and August are the months with the highest number of total homicides.