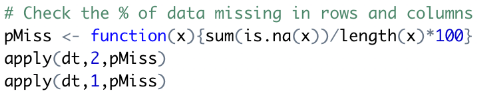
Programming Languages for Data Engineering
R Assignment
Julia Gratsova

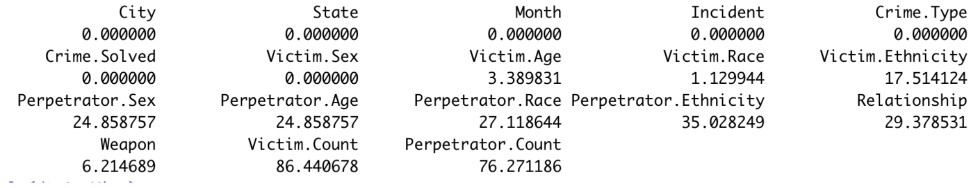
1. The particular dataset used here was derived from a much bigger dataset from Kaggle (www.kaggle.com/murderaccountability/homicide-reports/version/1#) which is called The Murder Accountability Project and is the most complete database of homicides in the United States currently available. For the purpose of this assignment, I selected only data for the year 1980, for 3 states: Hawaii, Idaho and Iowa. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used.

After loading the dataset into the data frame, I removed some of the variables, that were typically administration-related, such as type of the agency and also the Year variable, since all the data is from the same year. This dataset has a lot of unknown values, particularly from where the perpetrator has never been caught, but serves as an important indication of the unsolved crime rate and normally these wouldn't be replaced or removed. However, for the assignment I pre-processed the dataset, replacing all 'Unknown's and '0' values with NA and then checking the percentage of the missing data across the dataset using the following function:



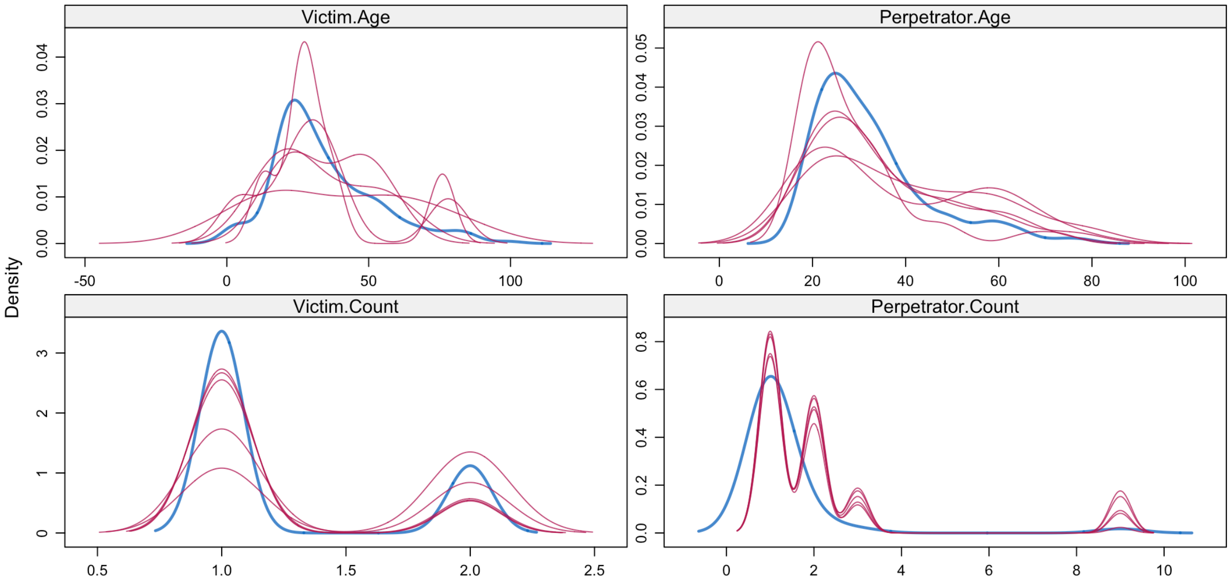
```
# Check the % of data missing in rows and columns
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(dt,2,pMiss)
apply(dt,1,pMiss)
```

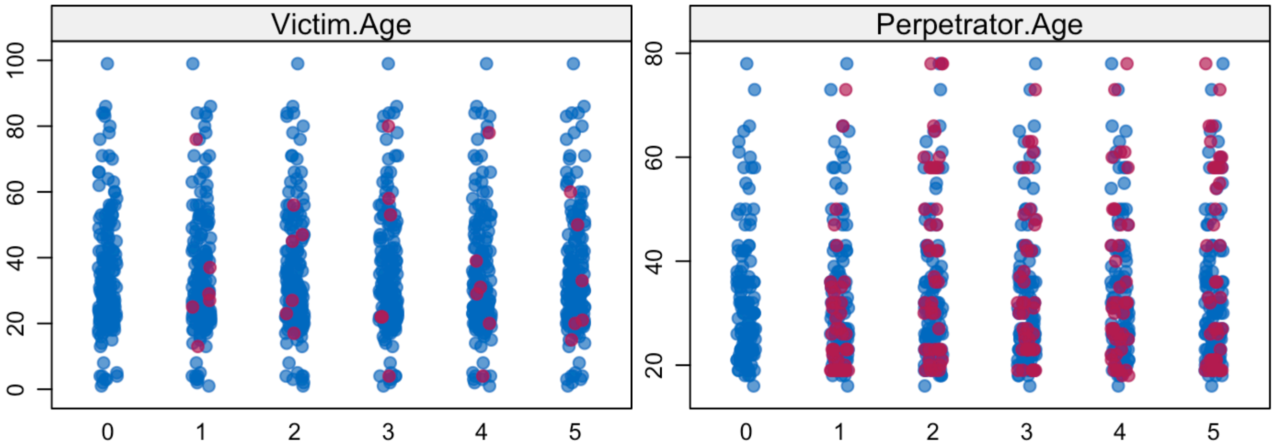
and receiving results, that indicate that some of the data concerning the Perpetrator has a lot of missing values and Victim and Perpetrator counts are unusable, missing over 86% and 76% of data respectively.

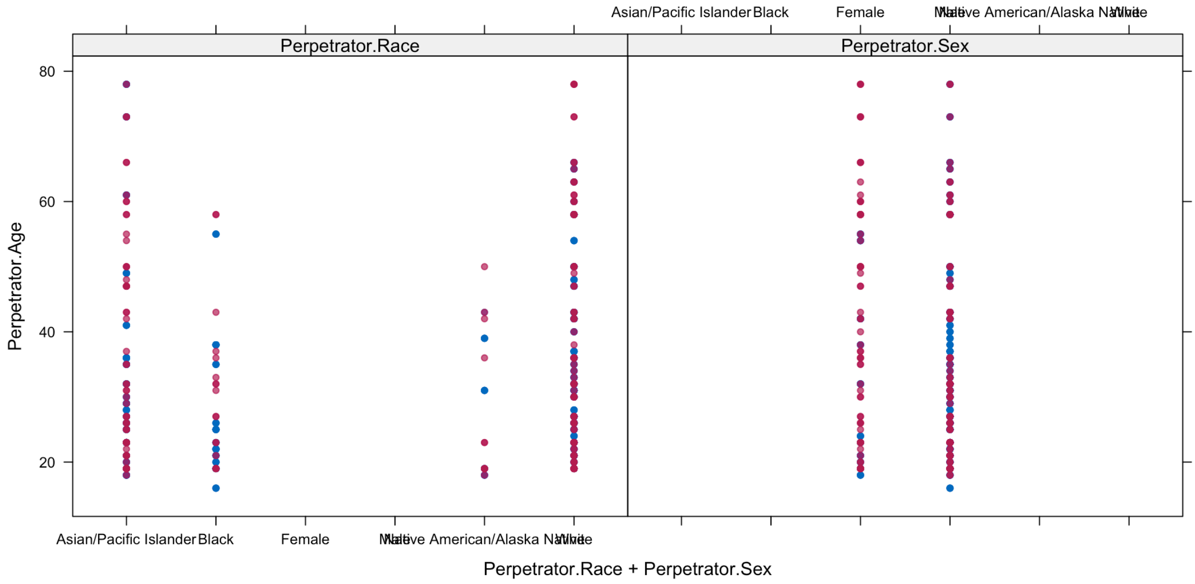


```
              City              State              Month           Incident         Crime.Type 
          0.000000           0.000000           0.000000           0.000000           0.000000 
      Crime.Solved         Victim.Sex         Victim.Age        Victim.Race   Victim.Ethnicity 
          0.000000           0.000000           3.389831           1.129944          17.514124 
   Perpetrator.Sex    Perpetrator.Age   Perpetrator.Race Perpetrator.Ethnicity     Relationship 
         24.858757          24.858757          27.118644          35.028249          29.378531 
            Weapon       Victim.Count  Perpetrator.Count 
          6.214689          86.440678          76.271186 
```

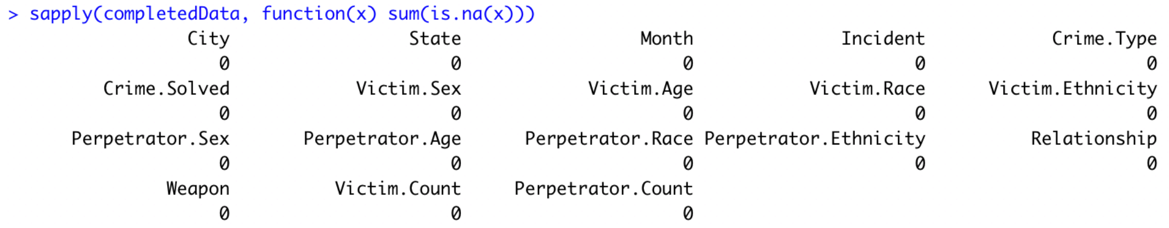
I decided to impute the data using the MICE package and then plot the distribution of the original and imputed data to inspect the results.





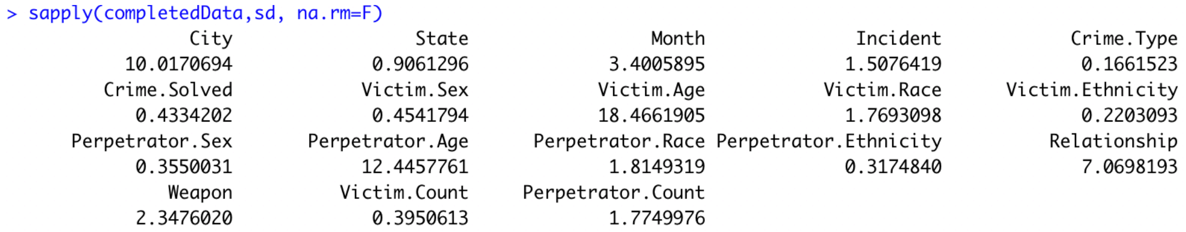


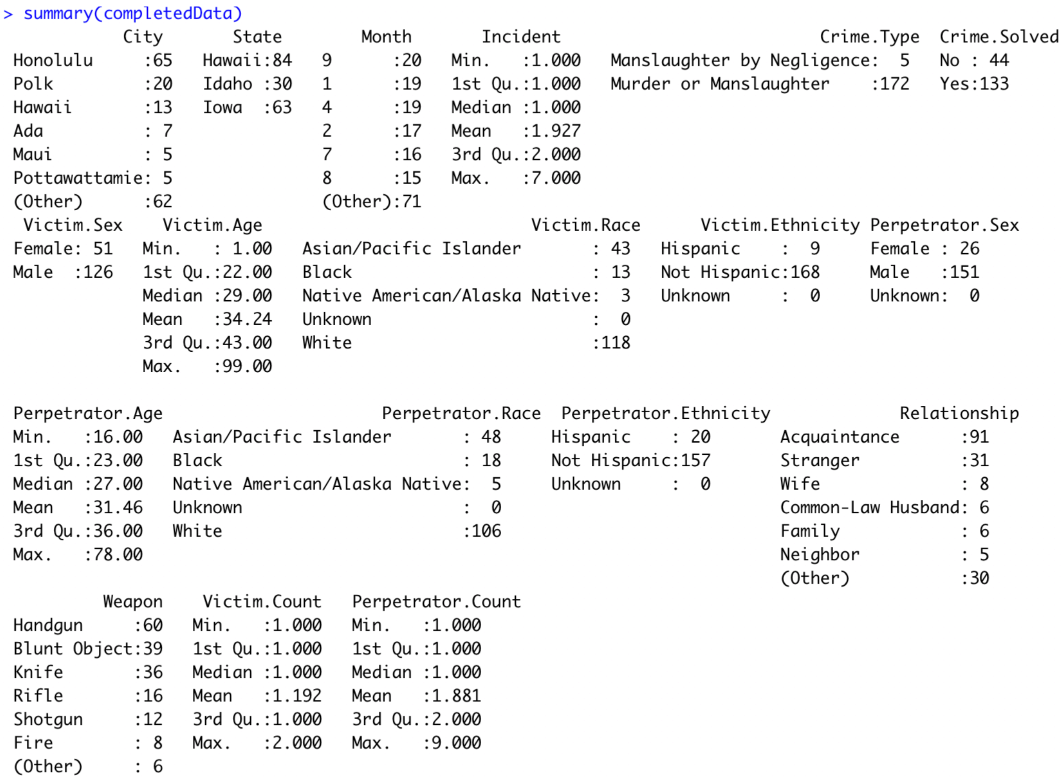
In the graphs above the imputed data shown in magenta and observed data in blue. We can see, that the distributions are very similar and generally imputed data follows the observed. I then checked if any missing data remained:


```
> sapply(completedData, function(x) sum(is.na(x)))
             City             State             Month          Incident        Crime.Type 
                0                 0                 0                 0                 0 
     Crime.Solved        Victim.Sex        Victim.Age       Victim.Race  Victim.Ethnicity 
                0                 0                 0                 0                 0 
  Perpetrator.Sex   Perpetrator.Age  Perpetrator.Race Perpetrator.Ethnicity      Relationship 
                0                 0                 0                 0                 0 
           Weapon      Victim.Count Perpetrator.Count 
                0                 0                 0 
```

As this dataset has a lot of categorical data, some of it was encoded into numerical for further analysis (Crime.Solved and Month).

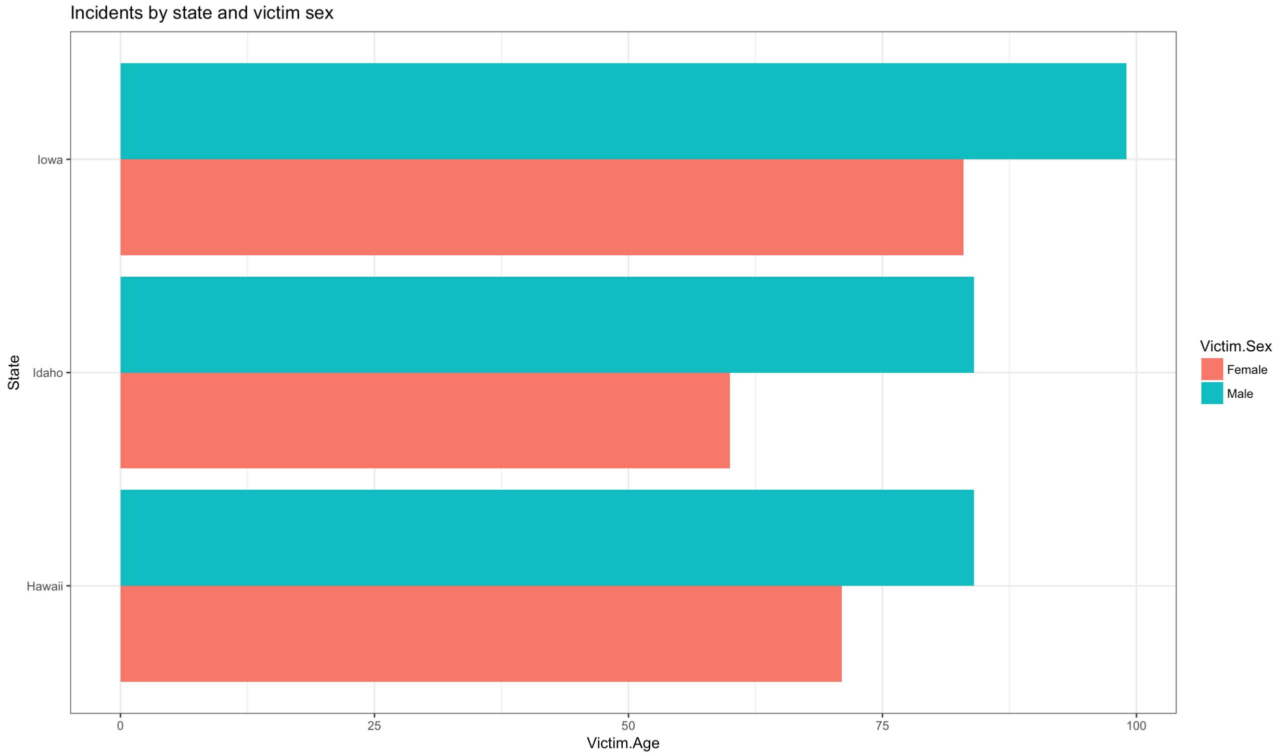
2. Basic statistic calculations were performed using summary() and sapply() to calculate the following: mean, median, 25th and 75th quartiles, min, max, sd.


```
> sapply(completedData,sd, na.rm=F)
             City             State             Month          Incident        Crime.Type 
       10.0170694         0.9061296         3.4005895         1.5076419         0.1661523 
     Crime.Solved        Victim.Sex        Victim.Age       Victim.Race  Victim.Ethnicity 
        0.4334202         0.4541794        18.4661905         1.7693098         0.2203093 
  Perpetrator.Sex   Perpetrator.Age  Perpetrator.Race Perpetrator.Ethnicity      Relationship 
        0.3550031        12.4457761         1.8149319         0.3174840         7.0698193 
           Weapon      Victim.Count Perpetrator.Count 
        2.3476020         0.3950613         1.7749976 
```

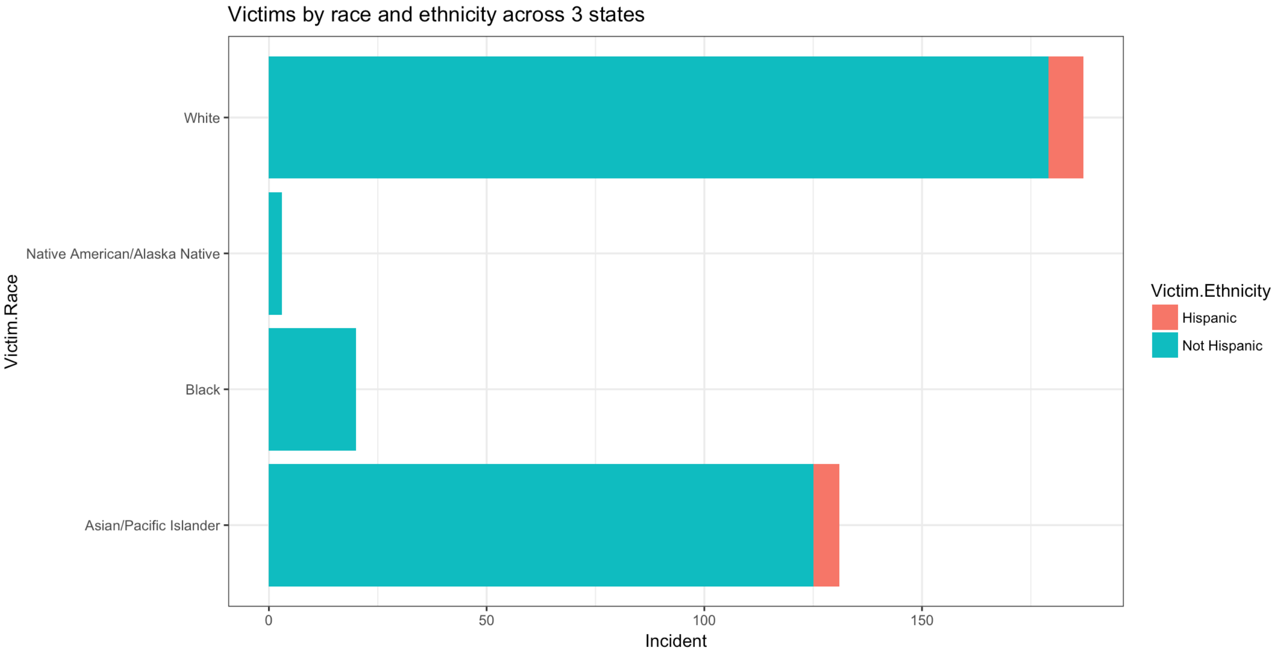


```
> summary(completedData)
           City          State        Month       Incident                            Crime.Type  Crime.Solved
 Honolulu     :65   Hawaii:84   9       :20   Min.   :1.000   Manslaughter by Negligence:  5   No : 44
 Polk         :20   Idaho :30   1       :19   1st Qu.:1.000   Murder or Manslaughter    :172   Yes:133
 Hawaii       :13   Iowa  :63   4       :19   Median :1.000
 Ada          : 7               2       :17   Mean   :1.927
 Maui         : 5               7       :16   3rd Qu.:2.000
 Pottawattamie: 5               8       :15   Max.   :7.000
 (Other)      :62               (Other):71
  Victim.Sex    Victim.Age                             Victim.Race      Victim.Ethnicity Perpetrator.Sex
 Female: 51   Min.   : 1.00   Asian/Pacific Islander       : 43   Hispanic    :  9     Female : 26
 Male  :126   1st Qu.:22.00   Black                        : 13   Not Hispanic:168     Male   :151
              Median :29.00   Native American/Alaska Native:  3   Unknown     :  0     Unknown:  0
              Mean   :34.24   Unknown                      :  0
              3rd Qu.:43.00   White                        :118
              Max.   :99.00

 Perpetrator.Age                          Perpetrator.Race  Perpetrator.Ethnicity             Relationship
 Min.   :16.00   Asian/Pacific Islander       : 48    Hispanic    : 20       Acquaintance      :91
 1st Qu.:23.00   Black                        : 18    Not Hispanic:157       Stranger          :31
 Median :27.00   Native American/Alaska Native:  5    Unknown     :  0       Wife              : 8
 Mean   :31.46   Unknown                      :  0                           Common-Law Husband: 6
 3rd Qu.:36.00   White                        :106                           Family            : 6
 Max.   :78.00                                                               Neighbor          : 5
                                                                             (Other)           :30
         Weapon     Victim.Count    Perpetrator.Count
 Handgun     :60   Min.   :1.000   Min.   :1.000
 Blunt Object:39   1st Qu.:1.000   1st Qu.:1.000
 Knife       :36   Median :1.000   Median :1.000
 Rifle       :16   Mean   :1.192   Mean   :1.881
 Shotgun     :12   3rd Qu.:1.000   3rd Qu.:2.000
 Fire        : 8   Max.   :2.000   Max.   :9.000
 (Other)     : 6
```

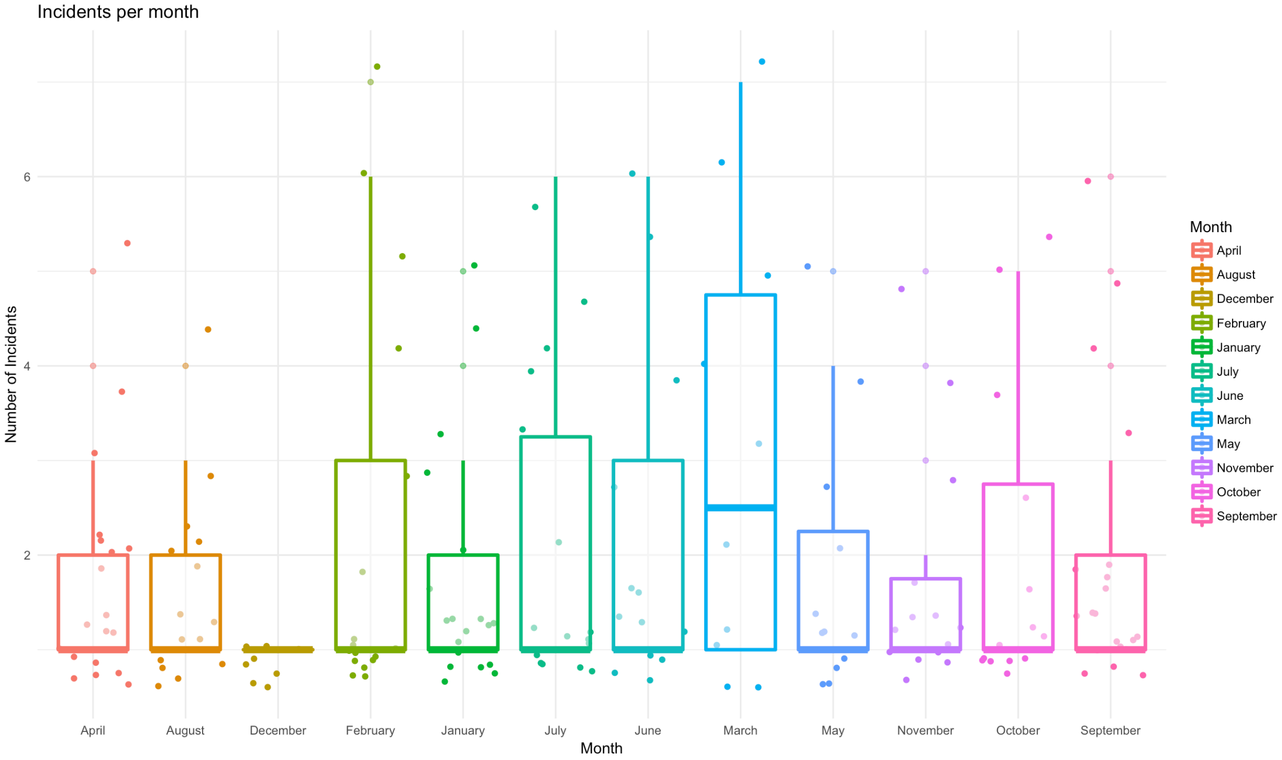
The data was then plotted in order to inspect the relationships in data:



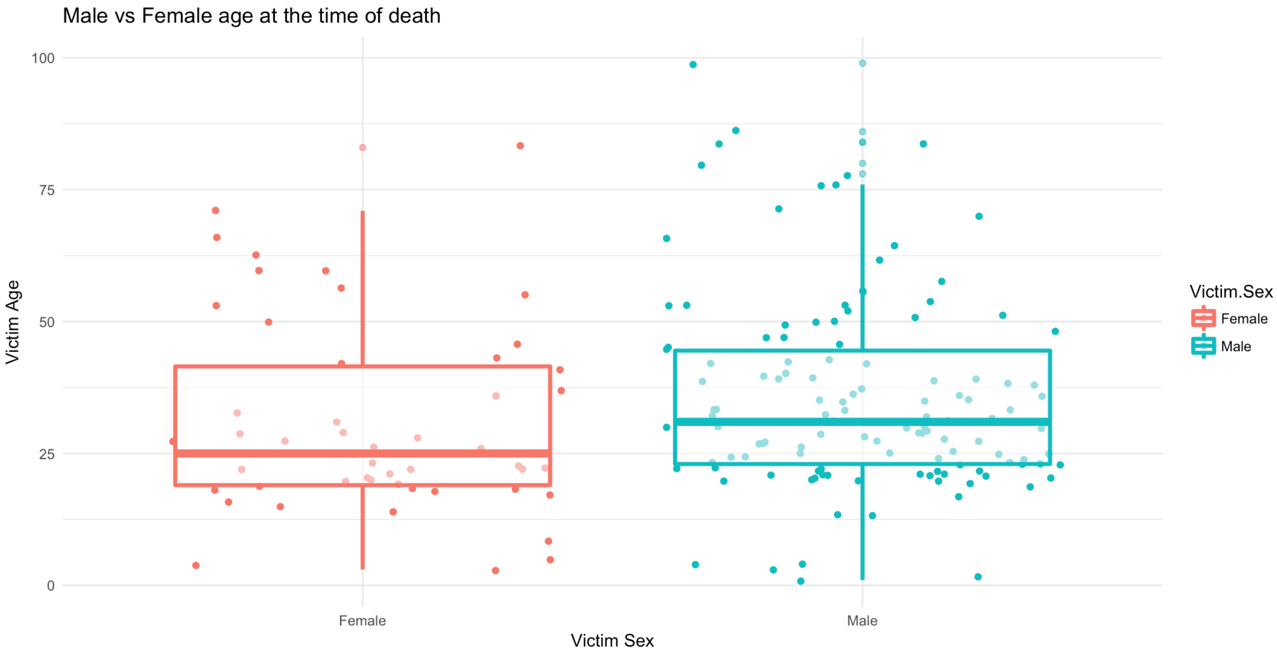
It can be seen, that males become victims more often than females, regardless on state.



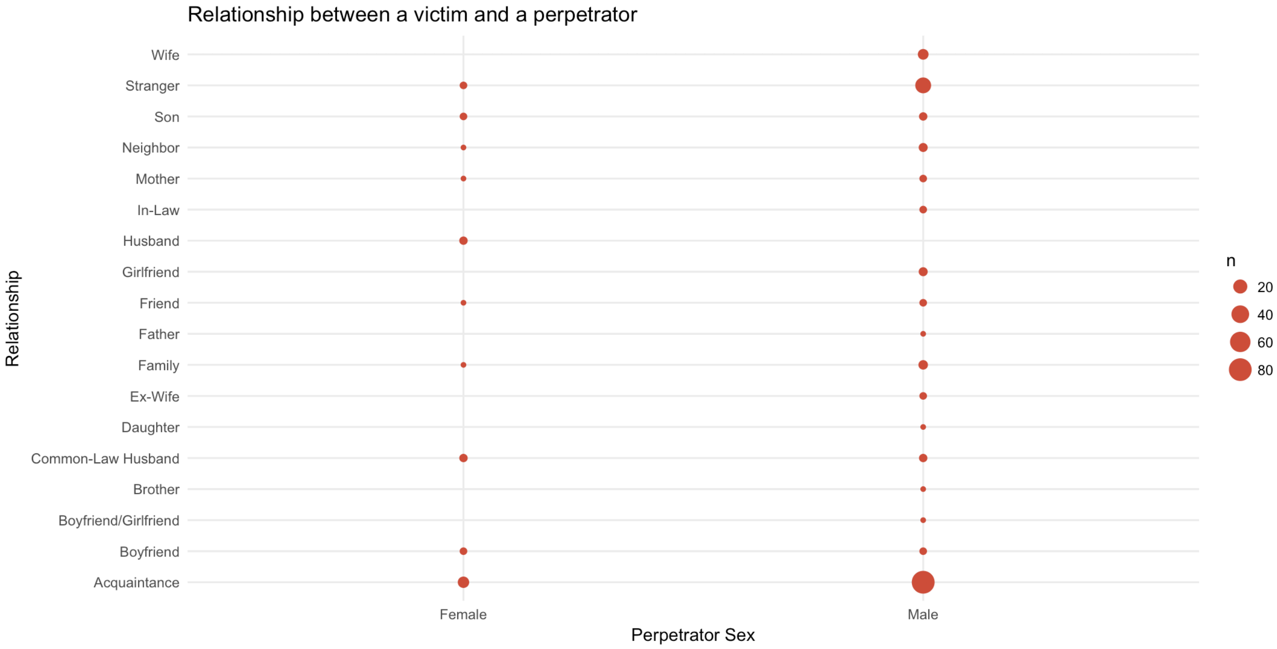
The high number of Asian/Pacific Islander victims is explained by the majority of data representing the Hawaii island.



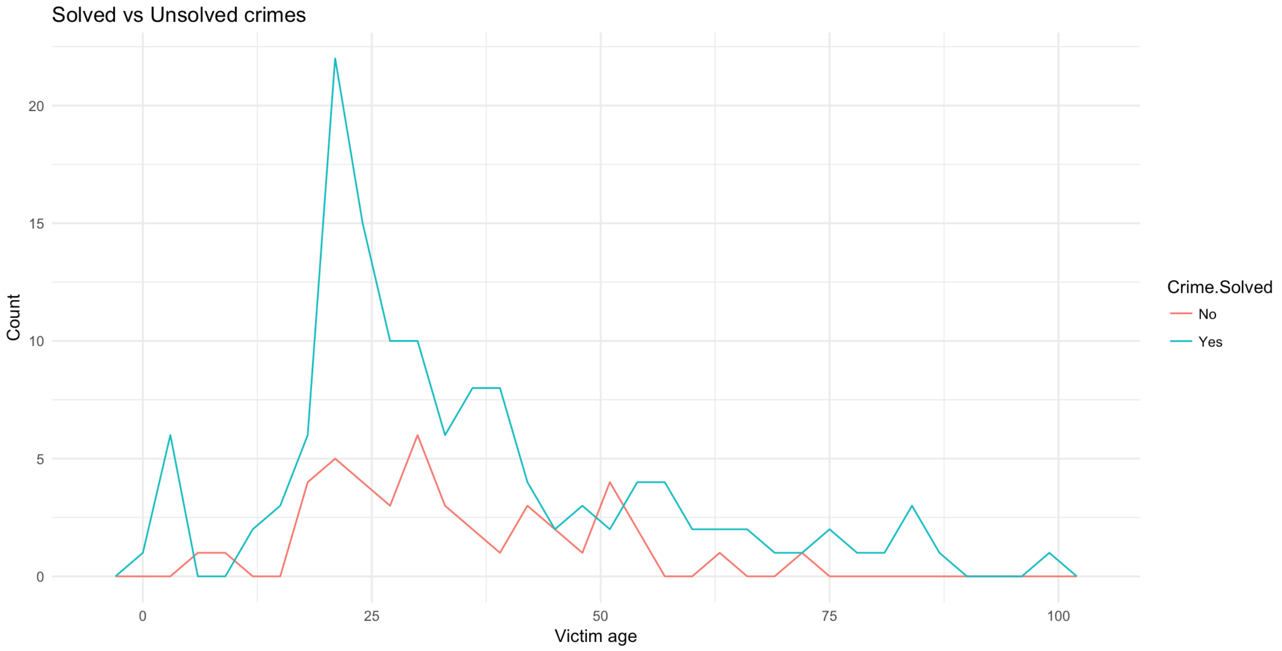
Although this box plot doesn’t represent a clear mean due to a large number of single incidents, it can be clearly seen that March has the most incidents overall, while December has the least body count.



This box plot shows that females on average are killed at a younger age than males.

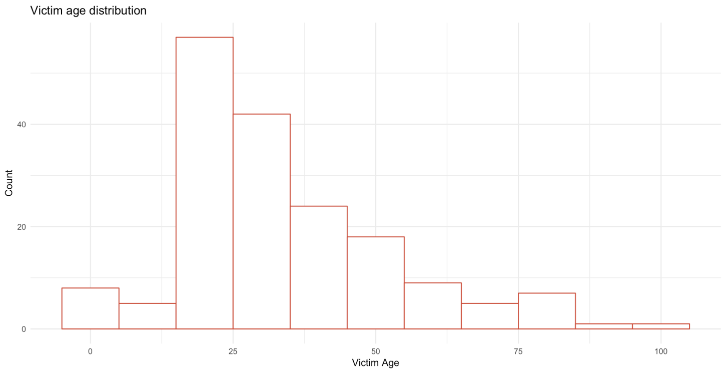


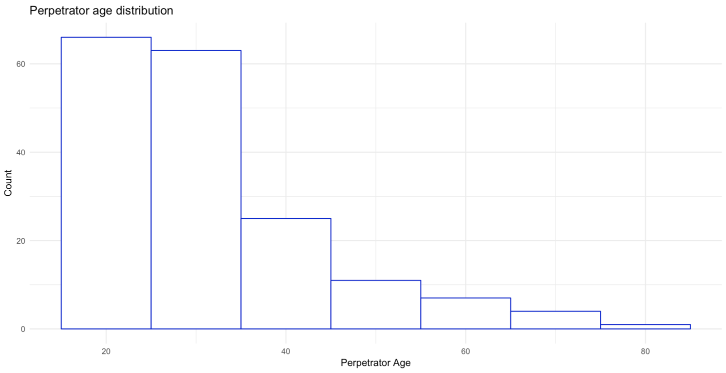
This plot is interesting, as it clearly states how much more often males are the Perpetrators than females, with the majority of the victims being merely acquaintances or strangers. While within the family, males are likely to go after any family member, females did not kill fathers, brothers or daughters.

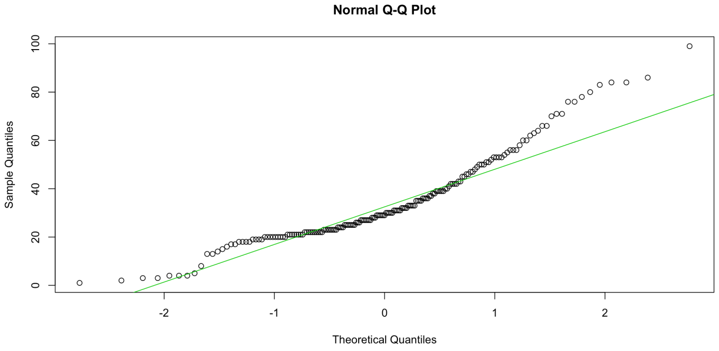


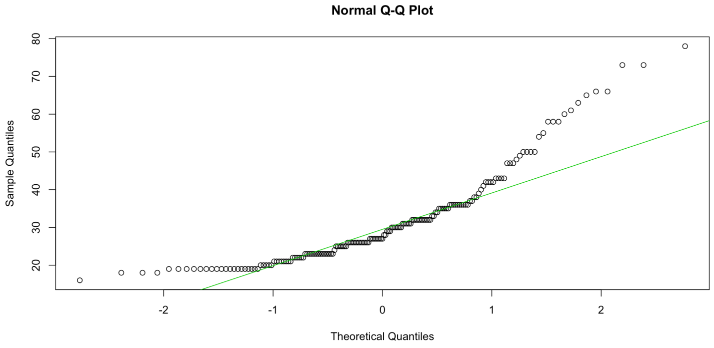
There is a spike of unsolved crimes for victims aged around 52 years old.

3. In order to check the data for normality, histograms were plotted along with density plots.



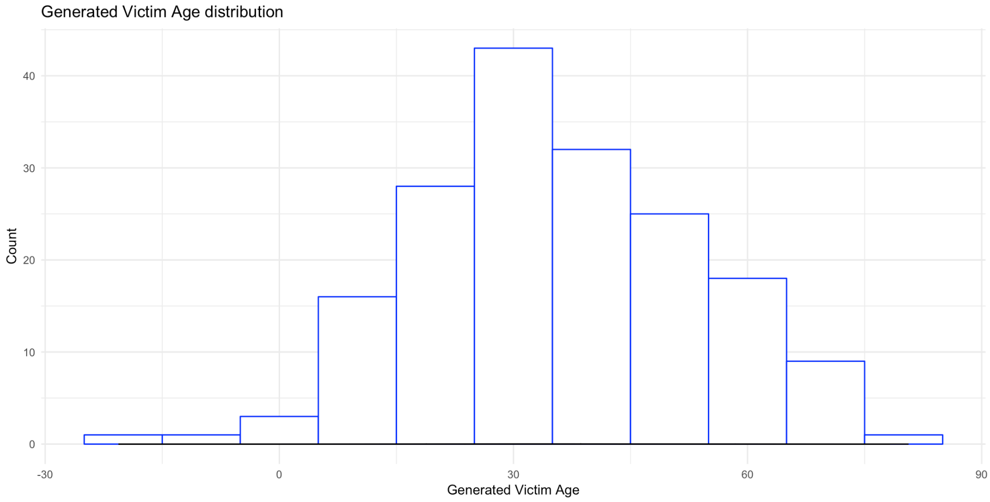


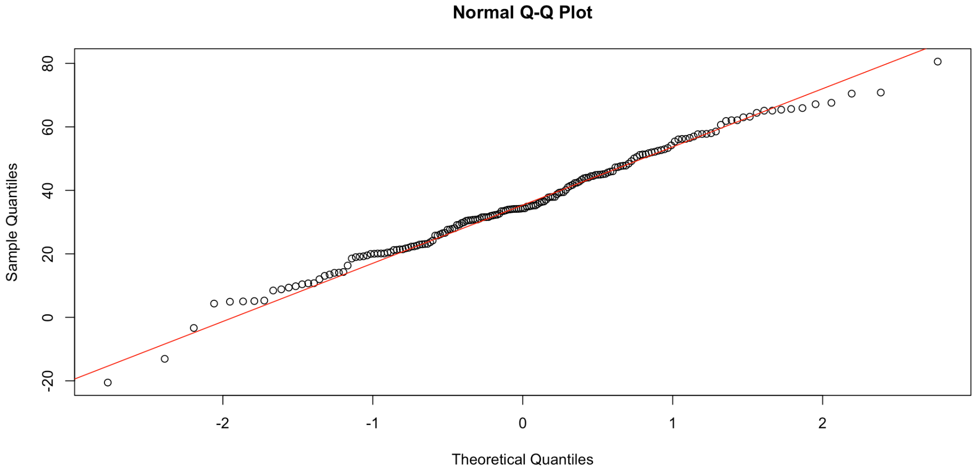




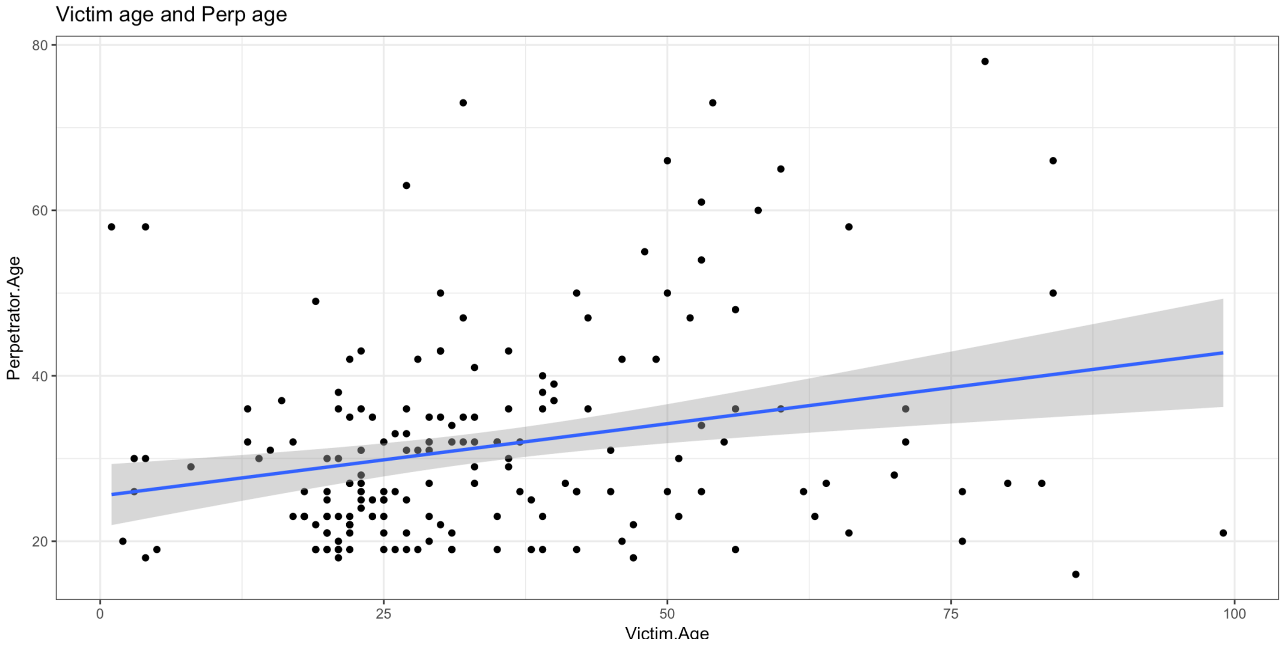
A random normal distribution was generated with the mean and standard deviation calculated earlier for Victim.Age in order to compare to the distribution represented by data.

It can be observed that the original data represented by Victim.Age is heavily skewed and the tails on its density plot lie outwit the line. We can conclude that original data doesn’t follow a normal distribution.





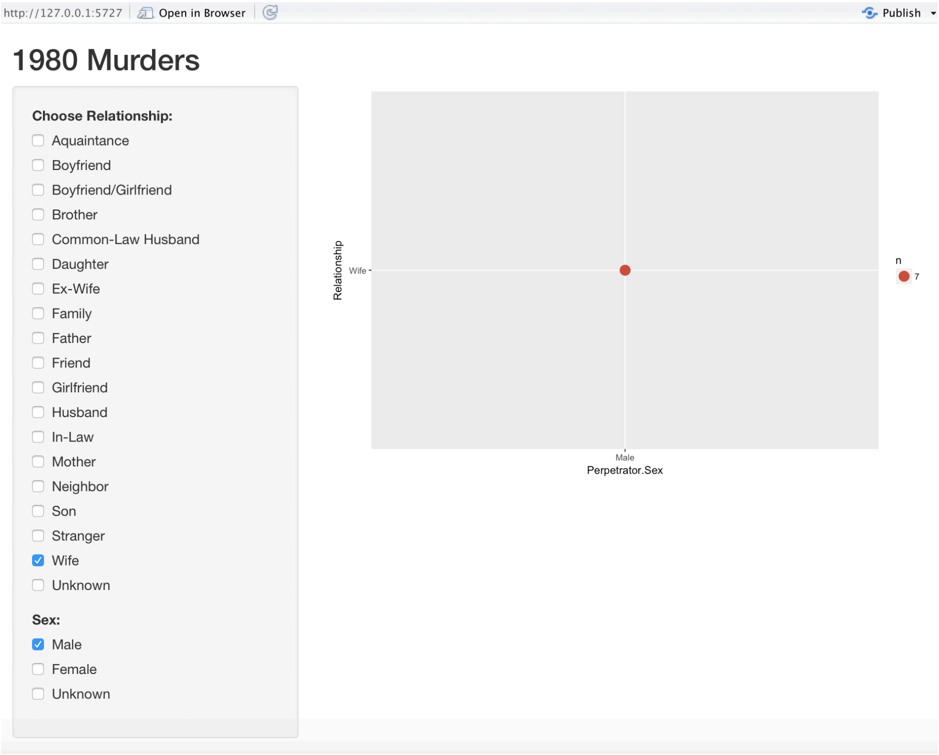
4. Linear regression was performed on some of the variables. Due to the much of the data being categorical, linear regression is not the best choice in the case of most of this dataset and logistic regression can be used instead in order to gain further insights into data. However, to demonstrate the work of linear regression, the following plot was constructed in order to check for relationship of victim and perpetrator ages:



```
lm(formula = Victim.Age ~ Perpetrator.Age, data = completedData)

Residuals:
    Min      1Q  Median      3Q     Max 
-43.452 -10.213  -4.443   8.018  68.787 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      22.1308     3.6631   6.042 8.94e-09 ***
Perpetrator.Age   0.3849     0.1083   3.553  0.00049 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.89 on 175 degrees of freedom
Multiple R-squared:  0.06728,	Adjusted R-squared:  0.06195 
F-statistic: 12.62 on 1 and 175 DF,  p-value: 0.0004899
```

Looking at the linear regression we can see that there is is evidence of the relationship between the victim ages and perpetrator ages, as the calculated p-value is less than 0.05 (0.0004899), and it is also evident by the behaviour and angle of the line on the plot.

5. Shiny app was made to explore the dataset and show the relationship between the Perpetrator sex and relationships to victims. However, the filtering is not working properly and is something I have tried to improve, but had run out of time. I will attach the code for both ui.R and server.R.



1980 Murders

Choose Relationship:

☐ Aquaintance
☐ Boyfriend
☐ Boyfriend/Girlfriend
☐ Brother
☐ Common-Law Husband
☐ Daughter
☐ Ex-Wife
☐ Family
☐ Father
☐ Friend
☐ Girlfriend
☐ Husband
☐ In-Law
☐ Mother
☐ Neighbor
☐ Son
☐ Stranger
☑ Wife
☐ Unknown

Sex:

☑ Male
☐ Female
☐ Unknown