Case Study Report

Introduction

This case study uses a dataset provided by the Murder Accountability Project, which is a nonprofit organization dedicated to addressing unsolved homicides in the United States. Their website can be accessed here: http://www.murderdata.org. As murder rates increase, and the solve rates decrease, more needs to be done. Since 1980, over 256,000 murders have gone unsolved. The dataset contains information for both solved and unsolved homicides since 1976 from all 50 states. This study focuses on the solved murders so that we can gain insight to apply towards the unsolved cases. These insights may also be used in future homicides so that they don't become unsolved. As we work through this case study, we will discuss the processes used to refine the dataset, the analysis used to explore the dataset, model selection, and the conclusions reached.

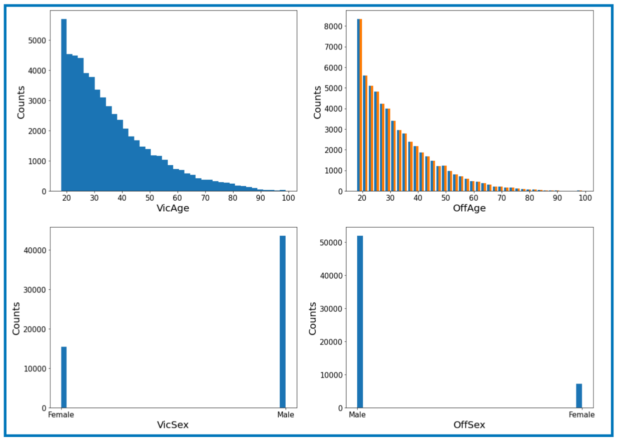
Process and Milestone Summaries

To remain focused on solved homicides between adults, crimes committed by/against anyone under the age of 18 were removed. Murders classified as "homicide by negligence" were also eliminated so that we could study homicides committed with intent. To keep the dataset manageable, additional reductions were made with only the columns and states of interest maintained.

- Columns: State, Year, Victim Age, Victim Sex, Victim Race, Offender Age, Offender Sex, Offender Race, Weapon, and Relationship.
- States: North Dakota, South Dakota, Nebraska, Kansas, Missouri, Wisconsin, Iowa, Illinois, Indiana, Michigan, and Ohio.

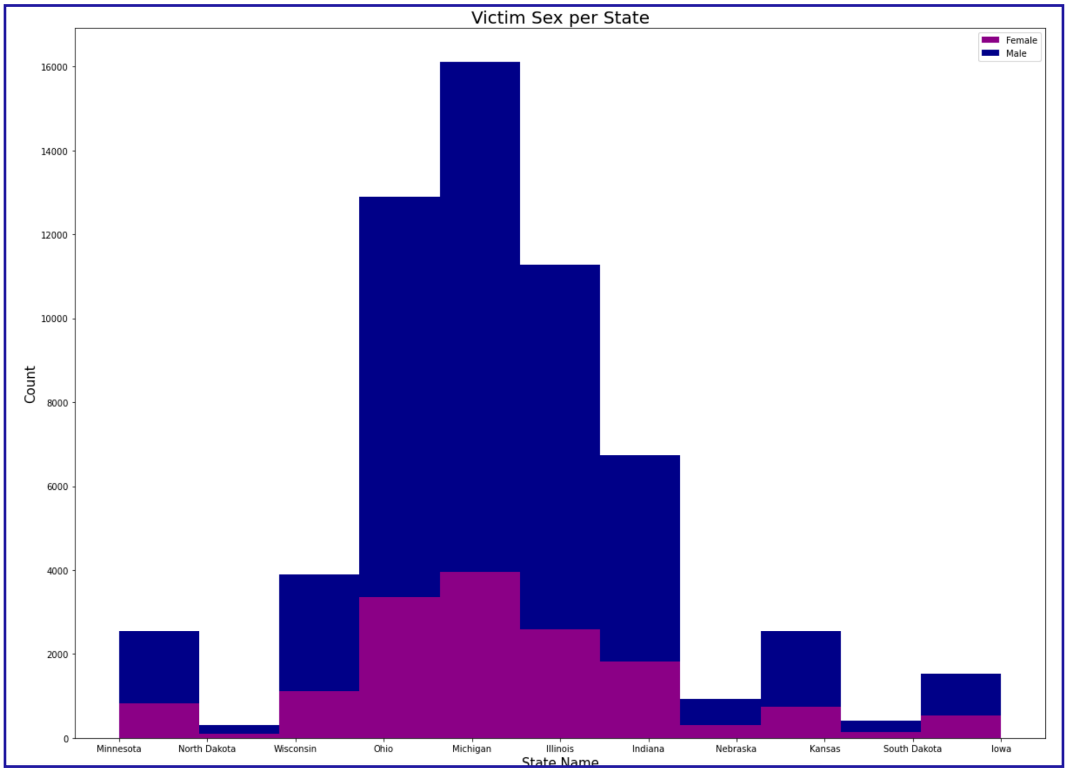
The first set of variables looked at were victim/offender sex, race, and ages, and this remained the area of interest until the full focus was centered on the victim's sex. Homicide rates were much higher among younger male adults than females. This is apparent in figure 1.

Figure 1. Victim/Offender Age & Sex



The next question to ask is whether or not this was true across all states. Figure 2 shows the comparison of victim sex by midwest state. Again, male victims numbered much higher than female victims. With the dark blue representing males, and the light purple representing females, it appears this gender gap is greatest in states with the highest homicide rates of Michigan, Ohio, and Illinois.

Figure 2. Victim Sex by State



Additional research was conducted by examining the victims sex according to age. The same trend held true that males were more often the victims than females, especially the younger they were. As both sexes age, this gender gap continues to get smaller. However, it never disappears, with males continuing to be more likely to be the victim of a homicide than females. This is shown in figure 3.

At this point in the case study I was still interested in what role, if any, race and gender played. According to the statistics

Figure 3. Victim age by Sex

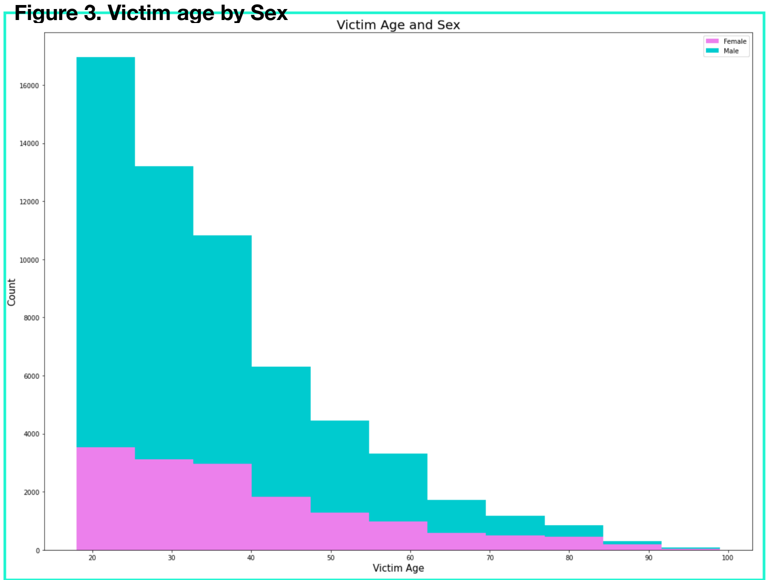
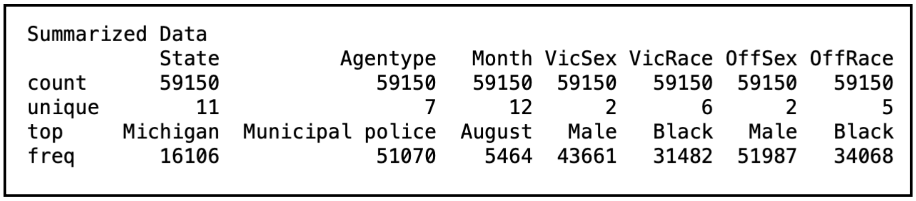


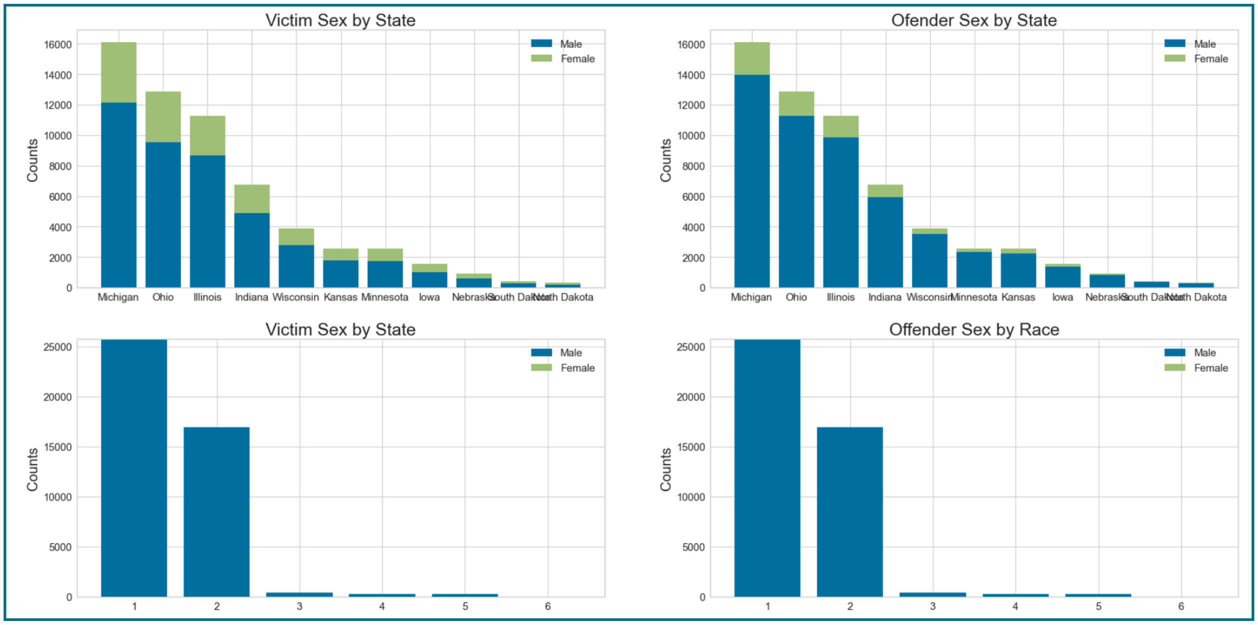
Figure 4. Summarized Data



```
Summarized Data
          State          Agentype   Month VicSex VicRace OffSex OffRace
count     59150             59150   59150  59150   59150  59150   59150
unique       11                 7      12      2       6      2       5
top    Michigan  Municipal police  August   Male   Black   Male   Black
freq      16106             51070    5464  43661   31482  51987   34068
```

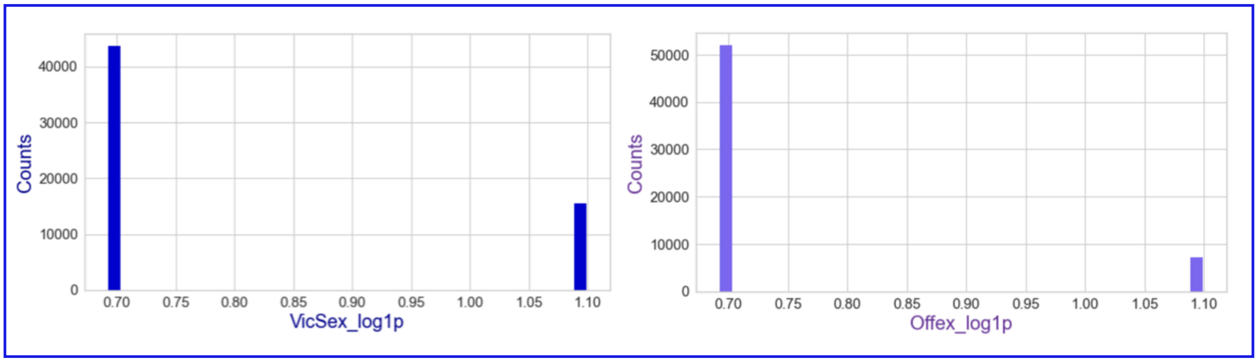
summary shown in figure 4, black males were most likely to be the victim of black males, with Michigan having the highest rate. This led to the examination of victim/offender sex and race showing in figure 5. To look at this further, I created the stacked bar charts to look at these categories.

Figure 5. Victim/Offender Sex and race



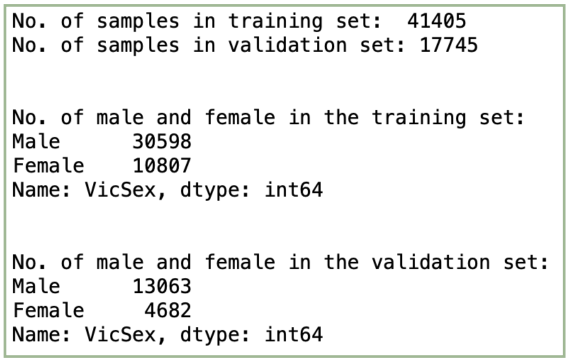
The case study is now becoming more focused on victim/offender sex and race. In order to remove some of the skewness of these variables, I did log transformations and corresponding histograms for victim and offender sex. As figure 6 shows, there is still a wide gap between genders for both victims and offenders.

Figure 6. Log transformations for victim/offender sex.



Now that the case study has worked through the possible variables, the focus has shifted to victim and offender sex and the question, “Can we use modeling to predict the victim and offender sex?” To do this, training and validation sets were created (figure 7). Once that was done, a confusion matrix was created for both victim sex and offender sex.

Figure 7. Training and validation sets for victim



```
No. of samples in training set:  41405
No. of samples in validation set: 17745


No. of male and female in the training set:
Male      30598
Female    10807
Name: VicSex, dtype: int64


No. of male and female in the validation set:
Male      13063
Female     4682
Name: VicSex, dtype: int64
```

For these models, the focus remained on victims and offenders in midwestern states. The first set in figure 8 looks at whether or not the victim was female or male. At the top of the figure is the confusion matrix to identify expected and actual results from the model. This will identify which values were correctly predicted and identify results that were true positive, false positive, true negative, and false negative. The bottom image is the logistic regression classification report to prove the f1 score. Precision tells us how many positives actually are positive, while recall tell us the number of positives of all positive examples. The f1 score measures the precision and recall to give us the final indicator of how good of a fit our model is. With a score of 1, our model is a very good fit.

Figure 8. Victim training and validation

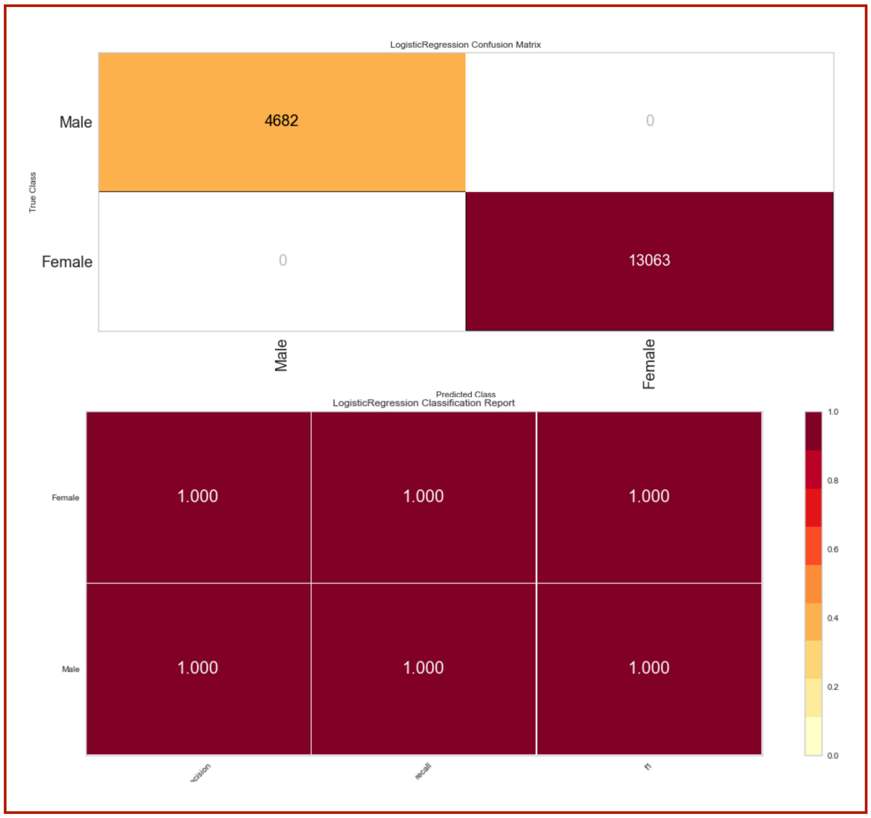
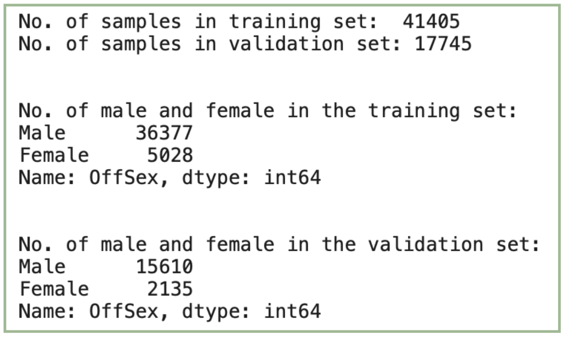


Figure 9 provides the training and validation data for the offender variable.

Figure 9 Training and validation sets for offender



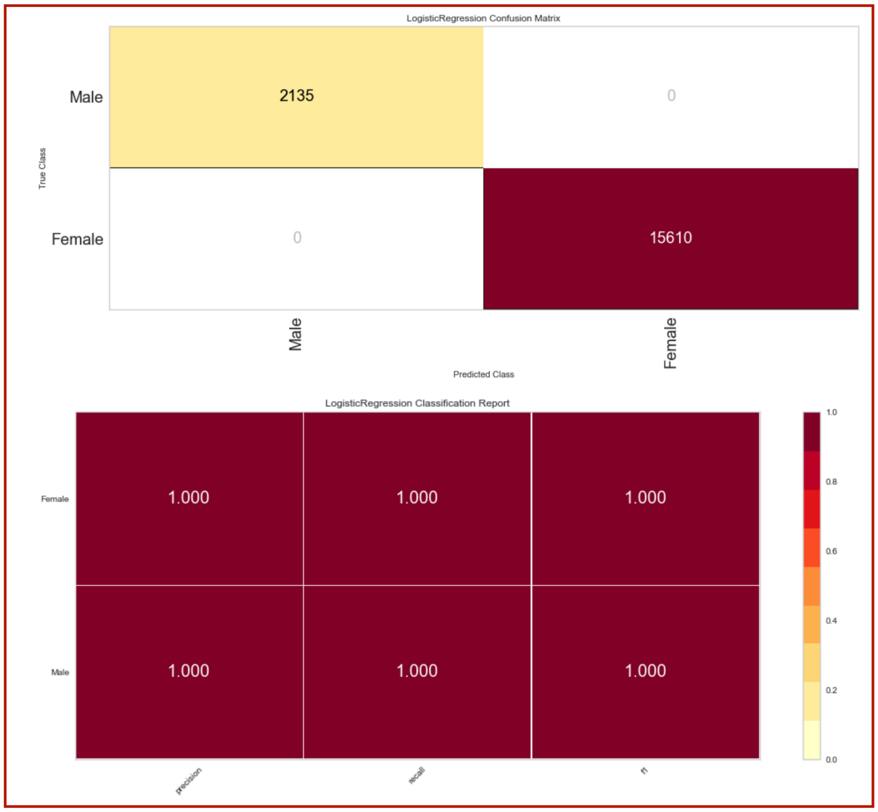
```
No. of samples in training set:  41405
No. of samples in validation set: 17745


No. of male and female in the training set:
Male      36377
Female     5028
Name: OffSex, dtype: int64


No. of male and female in the validation set:
Male      15610
Female     2135
Name: OffSex, dtype: int64
```

Figure 10 gives us the same two graphics of confusion matrix and classification reports for the offender that we did for the victim. Like the results above, this model seems to be a very good fit.

Figure 10. Offender training and



Conclusion/Recommendations

Given the results of these models, it seems like we could predict the offender’s sex when we know the victims’s sex. A word of caution about these results though - the data his heavily skewed towards males, even after the regression. This needs to be kept in mind when looking at the victim and searching for an offender. The fact that there is so much skew in the data, could have affected the results. Further study is recommended before relying on these results. Potential research on how additional variables affect the victim/offender sex would be worthwhile.

* A note on potential bias: throughout this case study, focus has been placed on the victim first due to my belief that victims should have higher priority than offenders. I recognize this is a bias that could affect future studies; however, in this case, I don’t believe this bias has any affect on the model outcomes.