Homicides in NYC: When are they domestic?

An analysis of various factors associated with domestic homicide

R-Rated - Holly Ansel, Jiayi Liang, Matt Jogodnik, Kehan Zhang

Section 1 - Introduction

For our final project, we (Team R-Rated) will be analyzing public data provided by the New York Police Department regarding homicides in the years 2016 through 2019. Our research question is the following:

Which factors make a homicide more likely to be domestic/are associated with domestic homicides?

The homicide data we are using comes from New York City's .gov website: https://www1.nyc.gov/site/nypd/stats/reports-analysis/homicide.page, and is collected by a department of the NYPD called CompStat. New York City's police department launched CompStat in 1994, successfully lowering crime rates through the use of professional management, statistical analysis, and implementation. All CompStat data is recorded by the NYPD in the Uniform Crime Reporting format. The data is collected by CompStat because every time a homicide is committed and reported it goes into their records. Each entry in the data set represents one recorded homicide incident in New York City. The variables include the date, the precinct, the victim and perpetrator's age, race, and sex, along with weapon used, arrest status, and other various notes. These variables serve to give important information on each of the homicides.

Section 2 - Data analysis plan

Since our research focuses on attributes of the victim and how they are/are not associated with domestic homicides, our main outcome variable will be domestic, a logical (Yes/No) variable associated with whether a homicide was committed by a family member of the victim or otherwise qualifies as being domestic. Predictor variables will include details surrounding the homicide such as what precinct or borough it was committed in, the time of year, and victim/perp age, sex, and race. While analyzing the data, we compare the domestic homicide with the non-domestic ones to see the role each predictable variable plays.

Statistical Methods

To conduct our analysis we plan on using the following statistical methods:

- Data Visualization

  Data visualization can help us break down and compare domestic homicides among different factors like race, sex, and age. Additionally, we can use map visualizations to show the proportion of homicides that are domestic in each borough to see if location has an effect on a homicide being domestic, directly answering our research question.

- Statistical Hypothesis Testing

  Through the use of statistical hypothesis testing, we can see if the proportion of homicides that are domestic in NYC is significantly different from other populations, such as NY state or the entire United States. In order to perform these tests, we would need domestic proportions for these areas, however, this data should not be difficult to obtain. Furthermore, since our data comes from multiple years, we could compare and see if one year has a statistically significantly higher proportion of domestic homicides than another and use this in our narrative on domestic homicides in NYC.

- Classification

  Similar to linear regression, we intend to use logistic regression as a method of classification. These tools will be of use in answering our research question based on a hypothetical homicide victim we could predict the likelihood that they were a domestic homicide victim.

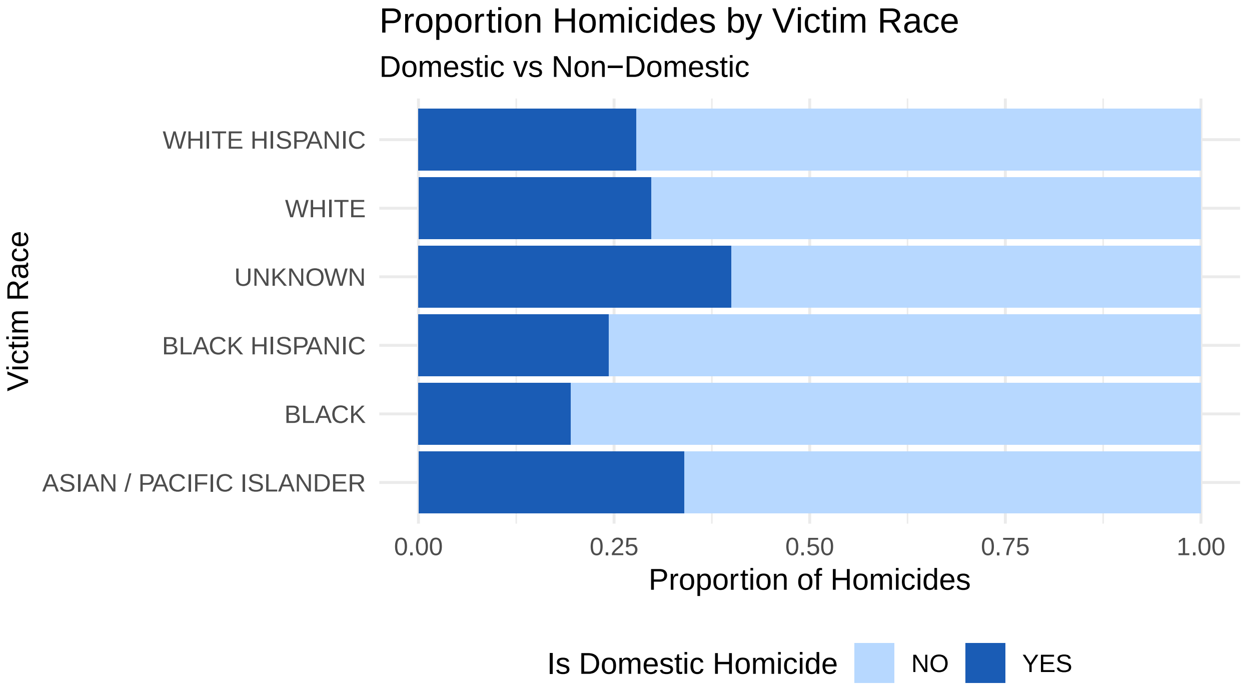
Visualizations

Below are sample visualizations that highlight simple qualities of the data that influence our analysis for answering the research question.

The majority of our summary statistics will be introduced in section 4 to connect the visualizations with our deeper analyses.

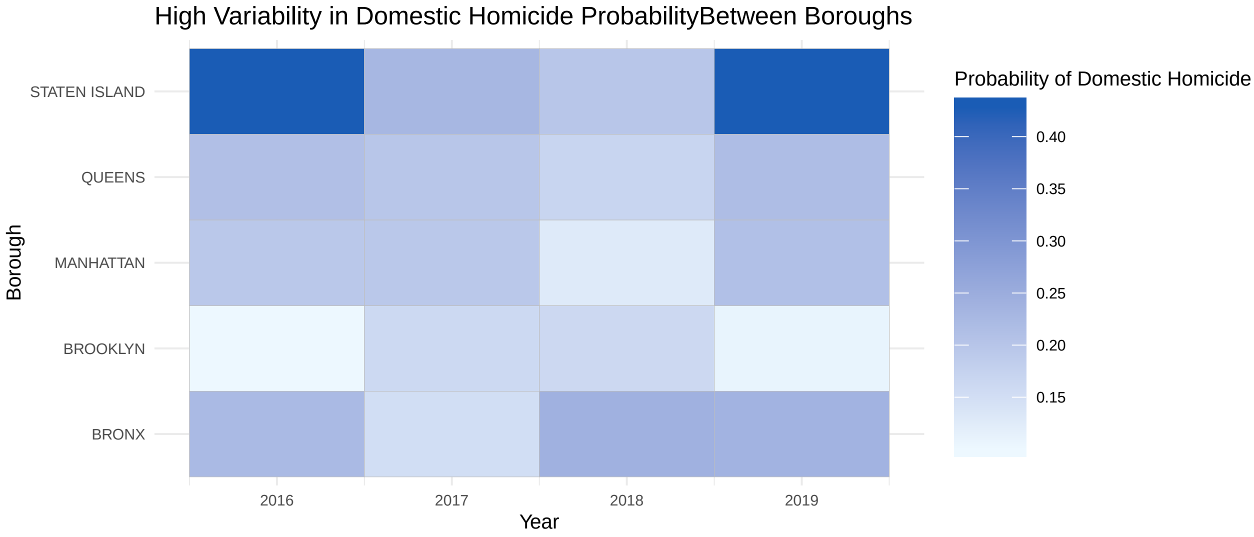
1. Domestic Homicides by Race Visualization

When looking at the proportion of domestic homicides by race it is evident that the proportion is highest among Asian/Pacific Islanders. This demonstrates how race can potentially contribute as a factor of what makes someone more likely to be a domestic homicide victim. This may be due to cultural factors that potentially influence familial relations or is coincidental and due to other circumstances.



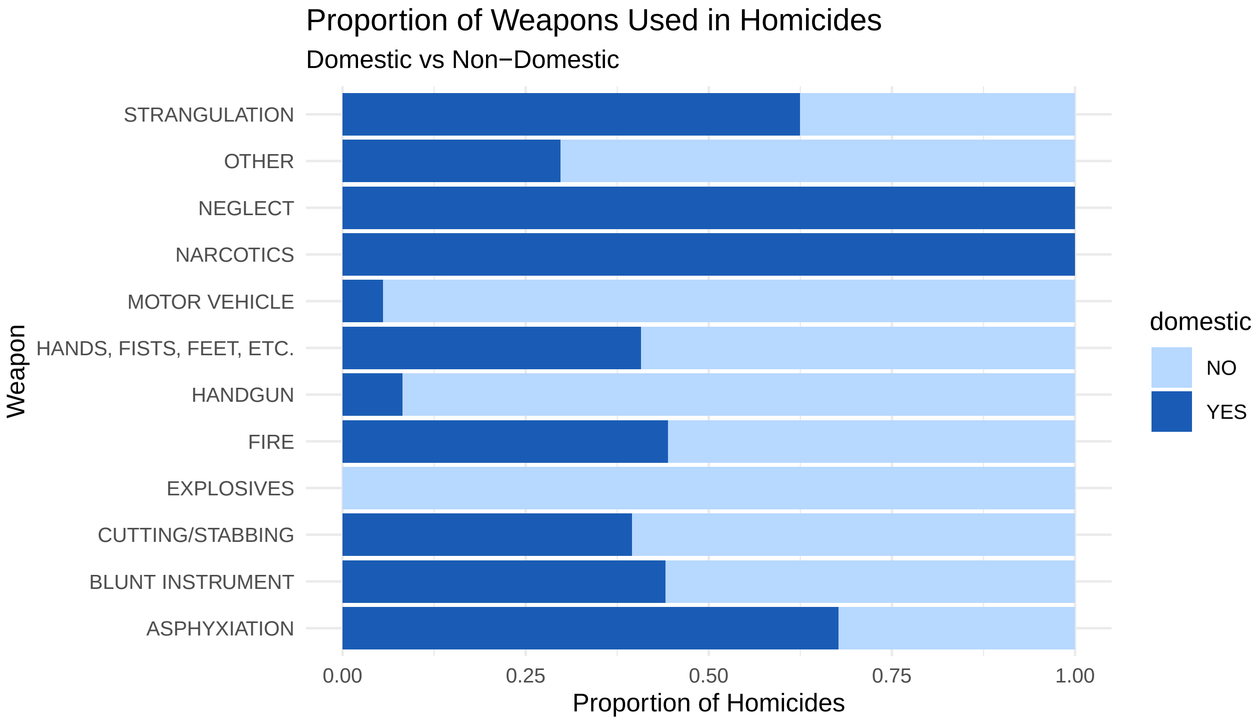
2. Domestic Homicide Probability By Borough Visualization

This visualization demonstrates the dispersion of domestic homicides across the boroughs of New York City each year from 2016 to 2019. Upon inspection, it is evident that Staten Island has the most domestic homicides out of the five boroughs. Additionally, it appears that Brooklyn has the least domestic homicides. It is plausible that this relationship could be related factors such as the average household income or environmental factors such as gang presence.



3. Proportion of Weapons Used in Domestic Homicides Compared To Regular Homicides Visualization and Domestic Homicide Weapons Used Summary Statistic

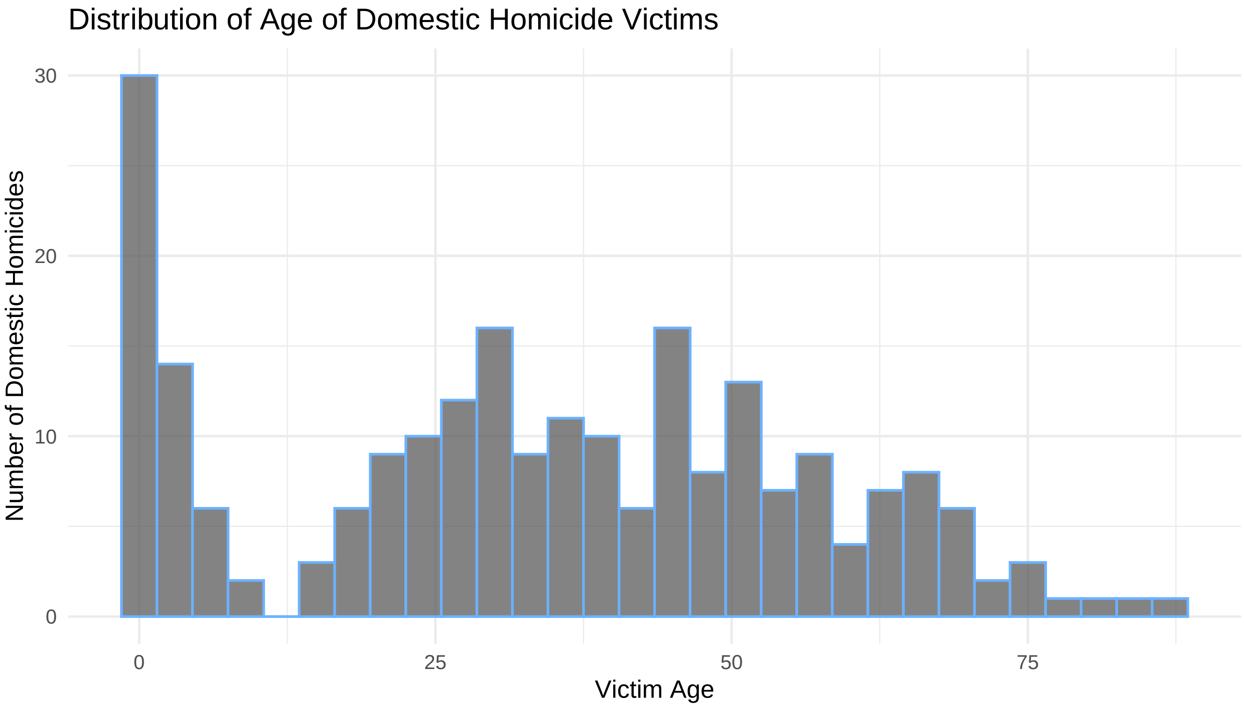
When examining the weapons used in domestic homicides, this visualization indicates that the weapons used for domestic and non-domestic homicides vary. Compared to regular homicides, domestic homicides use asphyxiation and strangulation proportionally more, as opposed to regular homicides which use handguns or explosives more.
The summary statistic shows that the incidents of neglect and narcotics, although they exclusively occurred in domestic homicides, were infrequent. When looking more at the summary statistic of the number of domestic homicides for each weapon type, it is evident that the most frequent weapons in domestic homicides is cutting/stabbing, handguns, and hands, fists, feet, etc. So although the proportions are lower compared to regular homicides, these are the most common weapon types for domestic homicides.



```
# A tibble: 11 x 2
# Groups:   weapon [11]
   weapon                     n
   <chr>                  <int>
 1 ASPHYXIATION              21
 2 BLUNT INSTRUMENT          15
 3 CUTTING/STABBING          93
 4 FIRE                       4
 5 HANDGUN                   43
 6 HANDS, FISTS, FEET, ETC.    31
 7 MOTOR VEHICLE              1
 8 NARCOTICS                  3
 9 NEGLECT                    1
10 OTHER                     14
11 STRANGULATION              5
```

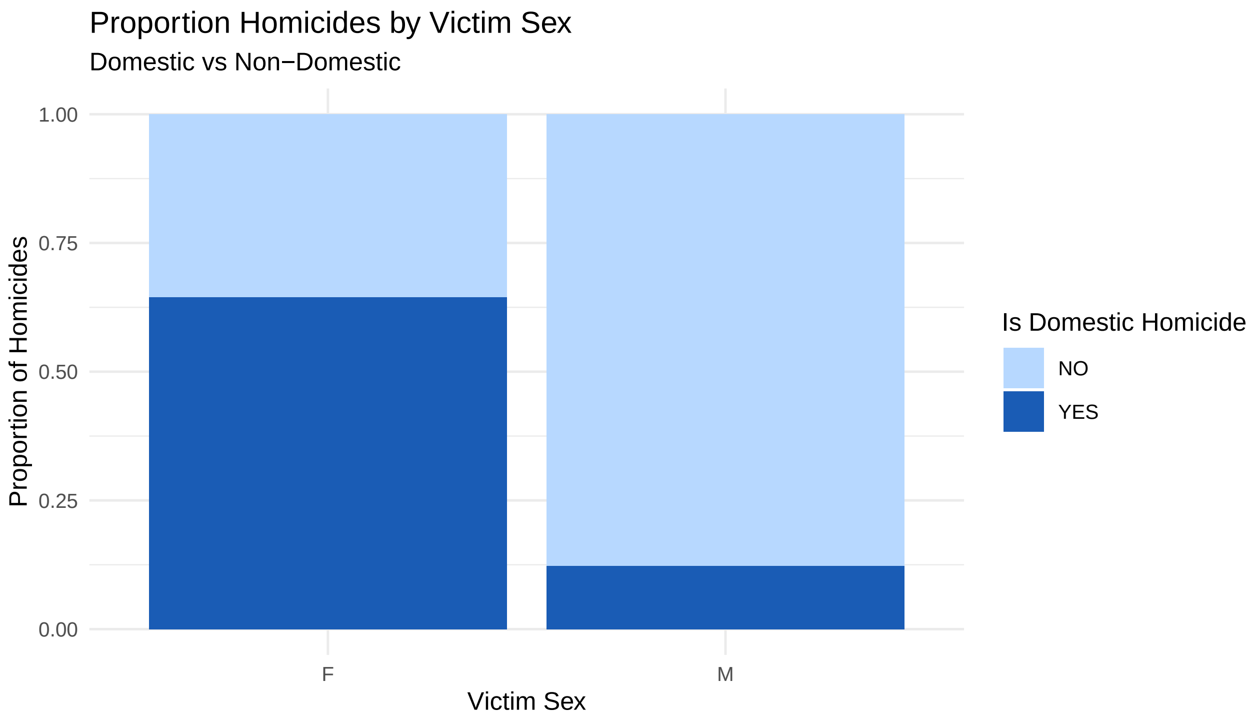
4. Distribution of the Age of Domestic Homicide Victims

The distribution of the age of domestic homicide victims illustrates that many victims are young children, especially infants. Among the adult population, the victim age is relatively evenly spread.



5. Domestic Homicides by Sex Visualization

In looking at the proportion of homicides by sex and distinguishing between domestic and non-domestic homicides it is evident that the proportion of females who are victims of domestic homicides is much higher than that of males.



Section 3 - Data

See data dimensions and code book in README.

```
Observations: 1,241
Variables: 28
$ shooting_homicide_incident_id_anony <dbl> 25126, 31093, 31236, 31245, 312...
$ date                                <date> 2019-04-27, 2019-07-03, 2019-0...
$ month                               <ord> Apr, Jul, Feb, Jan, Jan, Jan, J...
$ precinct                            <int> 41, 47, 75, 75, 33, 84, 43, 48,...
$ patrol_borough                      <chr> "PBBX", "PBBX", "PBBN", "PBBN",...
$ borough                             <chr> "BRONX", "BRONX", "BROOKLYN", "...
$ victim_age                          <dbl> 20, 25, 34, 29, 50, 44, 29, 38,...
$ victim_1                            <chr> NA, NA, NA, NA, NA, NA, NA, NA,...
$ victim_sex                          <chr> "M", "M", "M", "M", "F", "M", "...
$ victim_race                         <chr> "WHITE HISPANIC", "BLACK", "BLA...
$ victim_ethnic                       <chr> "HISPANIC", NA, NA, NA, "HISPAN...
$ perp_status                         <chr> NA, "ARRESTED", NA, NA, "DOA", ...
$ perp_age                            <dbl> NA, 34, NA, NA, 46, NA, 28, 38,...
$ perp_sex                            <chr> NA, "M", NA, NA, "M", NA, "M", ...
$ perp_race                           <chr> NA, "BLACK", NA, NA, "WHITE HIS...
$ perp_ethnic                         <chr> NA, NA, NA, NA, "HISPANIC", NA,...
$ relationship                        <chr> NA, NA, NA, NA, "INTIMATE PARTN...
$ weapon                              <chr> "HANDGUN", "HANDGUN", "HANDGUN"...
$ circumstance                        <chr> "UNKNOWN", "OTHER ARGUMENT", "G...
$ other_circumstance                  <chr> NA, NA, NA, NA, "DOMESTIC", NA,...
$ in_out                              <chr> "O", "O", "O", "I", "I", "I", "...
$ case_n                              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...
$ record_n                            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ victim_n                            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...
$ domestic                            <chr> NA, NA, NA, NA, "YES", NA, NA, ...
$ year                                <dbl> 2019, 2019, 2019, 2019, 2019, 2...
$ day                                 <int> 27, 3, 14, 1, 1, 4, 6, 7, 9, 10...
$ wday                                <ord> Sat, Wed, Thu, Tue, Tue, Fri, S...
```

Section 4 - Methods and Results

Before we do some further data analysis, we will first set the seed so our results are reproducible.

Next, we have to make some adjustments to the data-set to clean up our data. Since we will be running further tests on age, sex, and race (and obviously their relationship to domestic homicide), these are the 4 variables we will keep. Additionally, since we will be doing hypothesis testing, we select a small subset of our population (~20% = 200) so that our results can be compared with the entire population.

Proportion of Homicides that are Domestic

Using the simulation based method on a smaller sample of the data, we can generate confidence intervals for the proportion of homicides that are domestic. This statistic does a good job summarizing a key aspect of the data set that we are working with.

```
# A tibble: 1 x 2
  lower_bound upper_bound
        <dbl>       <dbl>
1        0.17       0.285
```

This confidence interval is generated with a simulation-based method, and indicates that we are 95% confident that the interval from 0.17 to 0.285 captures the true population proportion of homicides that are domestic in NYC from the years 2016 through 2019.

```
# A tibble: 1 x 2
  domestic prop_domestic
  <chr>            <dbl>
1 YES              0.235
```

As we can see by this sample statistic that looks at the entire data set, both statistics are extremely similar for the proportion of domestic homicides. The proportion of homicides in NYC that were committed by relatives of the victim or otherwise qualify as domestic between the years 2016-2019 was actually 23.45%, which lies within the confidence interval. This summary statistic demonstrates domestic homicides are not a rare event and that frequently homicide victims in NYC know and are very close to their killers. The similarity between our confidence interval and the summary statistic demonstrates that this sample which will be used in later parts of our analysis is representative of the population.

The Influence of Victim's Age

As seen in our visualization from part 2, age of victims of domestic homicide seems to be skewed very far to the left, suggesting people at very low age are disproportionately affected by domestic homicide. The summary table below confirms this, with approximately 92.7% of homicides involving a victim under 10 being domestic.

```
# A tibble: 7 x 3
# Groups:   victim_age_group [7]
  victim_age_group  prop total
  <chr>            <dbl> <int>
1 <10              0.927    55
```

```
2 >59               0.337    98
3 50-59             0.306    98
4 40-59             0.217   138
5 30-39             0.214   196
6 10-19             0.123    81
7 20-29             0.110   318
```

With this in mind, we wanted to examine the influence of victims' age on their likelihood of being a homicide victim so we performed an stimulation-based hypothesis test on the correlation between age and whether a homicide is domestic. For all the hypothesis tests in the project, we adopted the standard $\alpha = 0.05$.

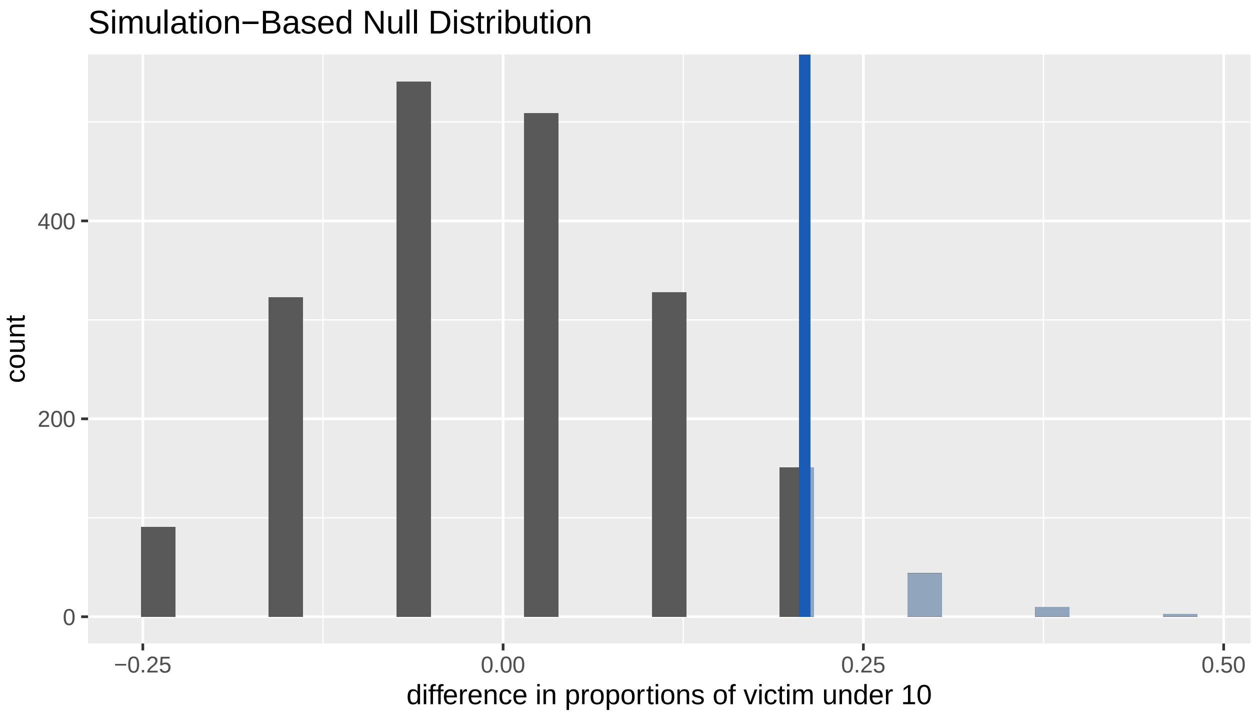
Hypotheses are as follows:

H_0 : Victim age and whether a homicide is domestic are independent.

$$p_{<10} = p_{\geq 10}$$

H_1 : Victim age and whether a homicide is domestic are not independent (in fact, people under the age of 10 are more likely to be the victim for domestic homicide).

$$p_{<10} > p_{\geq 10}$$



```
# A tibble: 1 x 1
  p_value
    <dbl>
1  0.0285
```

We find the p-value to be 0.029, which is less than the standard alpha of 0.05. Therefore, we can reject our null hypothesis that victim age and domestic homicide are independent. Extrapolated to the population as a whole, this poses the possibility (though does not definitively state) that victim age and the likelihood of a homicide being domestic could be linked.

The Influence of Victim’s Race

Furthermore, our visualization from part 2 shows that the race of victims of domestic homicide seems to be linked with a homicide being domestic, with Asian/Pacific Islanders showing the highest proportion of domestic homicide. The summary table below confirms this, with approximately 34% of homicides involving an Asian/Pacific Islander victim being domestic.

```
# A tibble: 5 x 3
# Groups:   victim_race [5]
  victim_race               prop total
  <chr>                    <dbl> <int>
1 ASIAN / PACIFIC ISLANDER 0.34     50
2 WHITE                    0.298    94
3 WHITE HISPANIC           0.279   201
4 BLACK HISPANIC           0.243    74
5 BLACK                    0.195   555
```

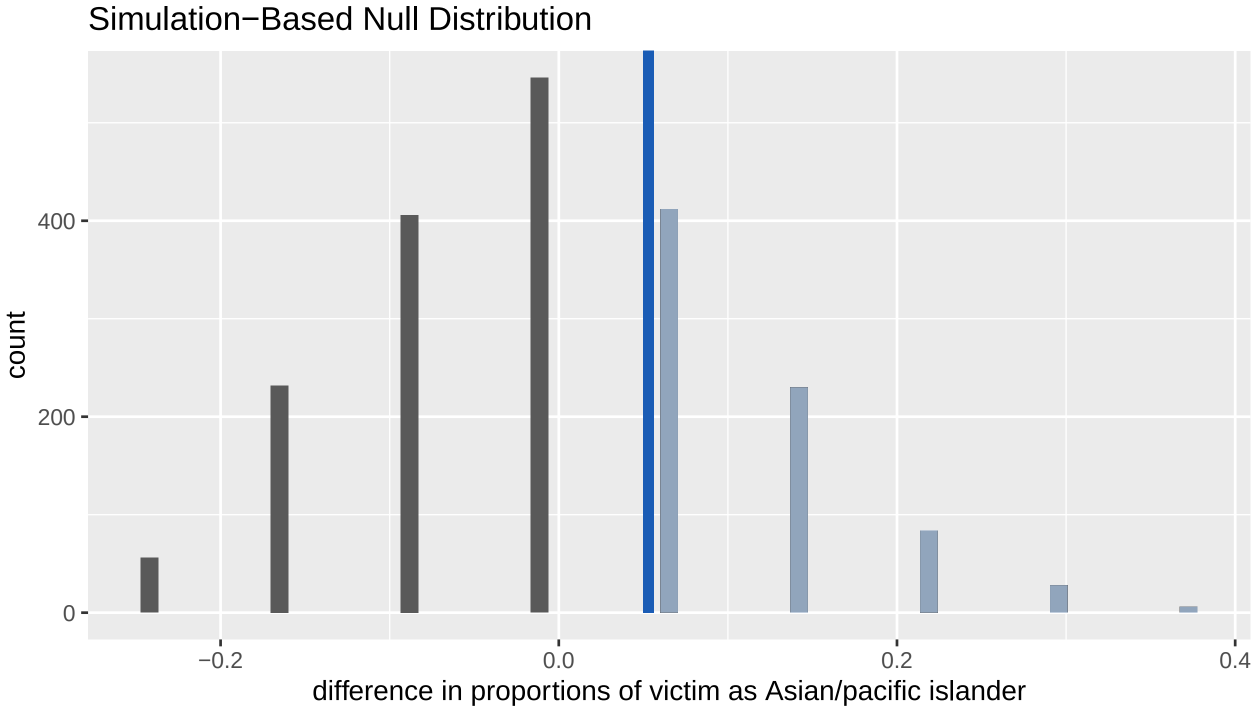
Seeing the proportions of different races as domestic homicide victims led us to perform a hypothesis test with the following hypotheses:

H_0 : Victim race and whether a homicide is domestic are independent.

$$p_{victim=asian/pi} = p_{victim \neq asian/pi}$$

H_1 : Victim race and whether a homicide is domestic are not independent (in fact, the proportion of domestic homicides among Asian/Pacific Islander victims is greater than among other races).

$$p_{victim=asian/pi} > p_{victim \neq asian/pi}$$



```
# A tibble: 1 x 1
  p_value
```

```
    <dbl>
1    0.38
```

We find the p-value to be 0.38, which is far greater than a standard alpha of 0.05. So we fail to reject our null hypothesis that victim race and domestic homicide are independent because we have insufficient evidence.

The Influence of Victim's Sex

Additionally the visualization relating to the sex of victims of domestic homicide from section 2 shows that female has a higher possibility of being the victim of domestic homicide. The summary table below confirms this, with approximately 64.5% of homicides involving a female being domestic compared to only 12.3% involving males being domestic.

```
# A tibble: 2 x 3
# Groups:   victim_sex [2]
  victim_sex  prop total
  <chr>      <dbl> <int>
1 F          0.645   211
2 M          0.123   772
```

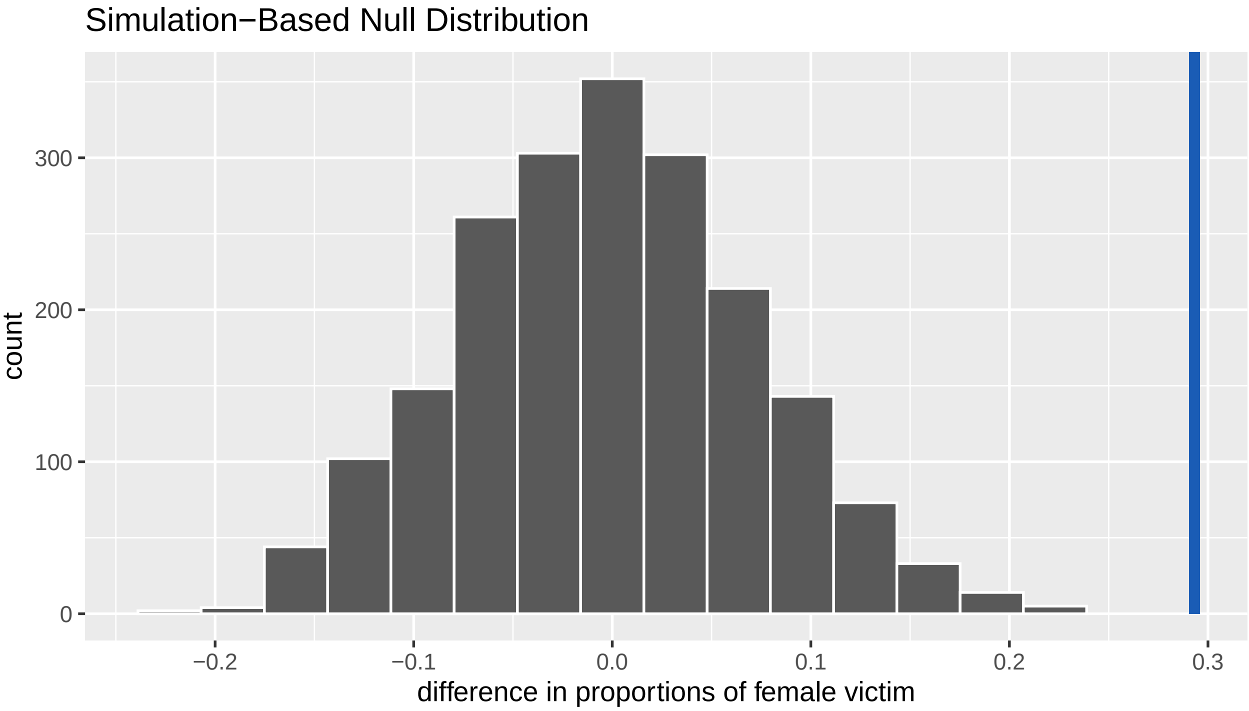
Similar to the previous variables analyzed, this led us to perform a hypothesis test with the following hypotheses:

H_0 : Victim sex and whether a homicide is domestic are independent.

$$p_{female} = p_{male}$$

H_1 : Victim sex and whether a homicide is domestic are not independent (in fact, the proportion of domestic homicides among female victims is greater than among male victims).

$$p_{female} > p_{male}$$



```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

We find the p-value to be 0, which being less than a standard alpha of 0.05 means that we can reject our null hypothesis that victim sex and domestic homicide are independent. Extrapolated to the population as a whole, this poses the possibility (though does not definitively state) that victim sex and the likelihood of a homicide being domestic could be linked.

Logistic Regression

Furthering our visualizations from part 2 on the effects borough, weapon, age, and sex have on the likelihood of a homicide being domestic, we will now perform logistic regression in an attempt to classify some observations in our data-set.

```
# A tibble: 17 x 5
   term                           estimate  std.error statistic  p.value
   <chr>                             <dbl>      <dbl>     <dbl>    <dbl>
 1 (Intercept)                      3.51      0.645      5.45   5.11e- 8
 2 weaponBLUNT INSTRUMENT          -1.08      0.705     -1.53   1.26e- 1
 3 weaponCUTTING/STABBING          -1.36      0.576     -2.37   1.80e- 2
 4 weaponFIRE                      -0.395     1.47      -0.268  7.89e- 1
 5 weaponHANDGUN                   -2.87      0.587     -4.90   9.61e- 7
 6 weaponHANDS, FISTS, FEET, ETC.  -1.05      0.638     -1.65   9.88e- 2
 7 weaponMOTOR VEHICLE             -3.65      1.22      -2.99   2.80e- 3
 8 weaponNARCOTICS                 13.9     734.         0.0190 9.85e- 1
 9 weaponNEGLECT                   14.6    1455.         0.0100 9.92e- 1
10 weaponOTHER                     -2.06      0.712     -2.89   3.81e- 3
11 weaponSTRANGULATION             -0.990     1.13      -0.873  3.83e- 1
12 victim_age                      -0.0170    0.00591   -2.88   3.96e- 3
13 victim_sexM                     -2.42      0.229    -10.6    3.48e-26
14 boroughBROOKLYN                 -0.350     0.268     -1.30   1.92e- 1
15 boroughMANHATTAN                -0.263     0.338     -0.778  4.37e- 1
16 boroughQUEENS                   -0.0163    0.304     -0.0537 9.57e- 1
17 boroughSTATEN ISLAND            -0.0776    0.454     -0.171  8.64e- 1
```

```
[1] 616.214
```

```
# A tibble: 7 x 5
  term                 estimate std.error statistic  p.value
  <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)            1.50     0.312       4.79  1.63e- 6
2 victim_age            -0.0112   0.00552    -2.02  4.32e- 2
3 victim_sexM           -2.51     0.210     -12.0   5.36e-33
4 boroughBROOKLYN       -0.458    0.253      -1.81  6.99e- 2
5 boroughMANHATTAN      -0.128    0.307      -0.418 6.76e- 1
6 boroughQUEENS          0.0294   0.280       0.105 9.16e- 1
7 boroughSTATEN ISLAND   0.136    0.409       0.333 7.39e- 1
```

```
[1] 675.6984
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   1.39     0.275        5.06 4.28e- 7
```

```
2 victim_age   -0.0119   0.00546     -2.17 2.98e- 2
3 victim_sexM  -2.52     0.209      -12.1  1.61e-33
```

```
[1] 672.4918
```

```
# A tibble: 28 x 5
   term                           estimate std.error statistic     p.value
   <chr>                             <dbl>     <dbl>     <dbl>       <dbl>
 1 (Intercept)                       3.49      0.657   5.32    0.000000105
 2 weaponBLUNT INSTRUMENT           -1.02      0.715  -1.43    0.153
 3 weaponCUTTING/STABBING           -1.32      0.587  -2.25    0.0246
 4 weaponFIRE                       -0.439     1.57   -0.280   0.779
 5 weaponHANDGUN                    -2.85      0.598  -4.77    0.00000180
 6 weaponHANDS, FISTS, FEET, ETC.   -0.971     0.649  -1.50    0.135
 7 weaponMOTOR VEHICLE              -3.78      1.28   -2.96    0.00309
 8 weaponNARCOTICS                  13.9     750.      0.0185  0.985
 9 weaponNEGLECT                    14.3    1455.      0.00979 0.992
10 weaponOTHER                      -1.96      0.720  -2.73    0.00638
# ... with 18 more rows
```

```
[1] 630.4961
```

We test 4 logistic regression models, using calculated AIC values to determine the best model. We find the first to have the lowest AIC, and thus choose this one for classification purposes.

```
[1] 0.78
```

We find that, given our first model, 78% of our test data was classified correctly. This figure suggests we have built a successful logistic regression model.

And, just to confirm that race is not a great factor for determining domestic homicide, we will incorporate it with our best model.

```
[1] 621.501
```

We find that AIC has increased by adding race, and thus, adding race worsens the fit of the model, agreeing with our hypothesis test above.

Section 5 - Discussion

For our final project, we focused on determining the relationship between various factors of homicide and the likelihood of a homicide being domestic. We conducted basic hypothesis tests of independence, bootstrapping intervals of confidence, and logistic regression to determine which of our data set's factors best predicted or were connected to domestic homicides. We conclude that domestic homicides are in fact influenced by many factors, among them and in loose proposed order of effect victim sex, victim age, weapon used, and borough. The extent to which these factors affect domestic homicide is subject for further analysis.

We started our analysis by conducting several preliminary visualizations which led us to select victim age, victim sex, victim race, weapon used, and borough as our factors to explore. Hypothesis testing for independence conducted on victim age and victim sex rejected our null hypotheses, suggesting a possible connection between these factors and domestic homicide. For victim sex, our analysis showed that females are much more likely to be domestic homicide victims than males. Similarly for age our hypothesis test concluded that young children (under the age of ten) are the most susceptible age group to be a victim of domestic homicide. Conversely, we failed to reject our null hypothesis for race, so we have insufficient evidence to make a conclusion about the role of race in domestic homicide victims. Logistic regression also supported a connection between domestic homicide and victim age and sex as the best model found incorporate these factors with decent (78%) success. Logistic regression also suggested a connection between borough and weapon used and domestic homicide, suggesting these factors may also be good predictors of domestic homicide. Logistic regression incorporating victim race increased its AIC, which was in agreement with results from the hypothesis test suggesting race to be a lesser factor in determining likelihood of domestic homicide.

In tying these conclusions back to our research question our current analysis presents females and children as two groups that are more likely to be victims of domestic homicide. Furthermore through the logistic regression we have also seen that the borough one lives in, such as Staten Island, may influence the likelihood that someone is a domestic homicide victim. However, our analysis indicated that we do not have enough evidence to make conclusions about the effect of the victim's race. The logistic regression also demonstrated that specific methods or weapons can serve as predictors for domestic homicide so the form of the homicide differs compared to that of a stranger. Together the analysis shows that the circumstances revolving around domestic homicides differ from those of regular homicides, and unfortunately, are not uncommon events in NYC.

Reflecting upon our methods, it would've been helpful to compare logistic regression classification results with those from k-NN, however, the fact that our data set was almost entirely comprised of categorical variables made this impossible as k-NN relies on distance. This also prevented us from developing linear models and using the step function to arrive at the best one – our original plan for classification modeling. Additionally, a chi-square test for independence could have tested all levels of a factor rather than just selecting the highest proportion level. However, we had difficulty in interpreting these results and thus stuck with our original analysis. This would have nonetheless made our hypothesis tests more reliable for what we were testing. Finally, it could have helped our analysis to make demographic comparisons to the complete NYC population. It is possible that disparities in domestic homicide proportion are high among certain groups simply because there are fewer observations and thus the data set isn't representative of this demographic. This is supported by both Asian/Pacific Islander and Victim <10 having only around 50 observations each. Further analysis of domestic homicide in NYC could address these concerns to arrive at a more detailed and probable conclusion.

Further analysis on this data set should focus on improving the reliability and level of detail of the conclusion. This would likely include starting out by creating more summary statistics to create more numerical variables which would allow for greater modeling/bootstrapping. Analyzing and investigating more variables would also create greater confidence in those selected. Comparisons along the way with actual population statistics from NYC would help put results and proportions in context with their demographic's relative representation. And analysis of multiple variables simultaneously such as seeing the difference between male and female children under the age of 10 as homicide victims would also give a deeper understanding. Additionally doing more analysis on whether there are multiple victims in one domestic homicide incident.Finally, chi-square testing would allow for better analysis of the relationship between each variable's several factor levels and domestic homicide. These changes would create more robust and generalizable results, leading to a better analysis of our research question: which factors of homicide increase its likelihood of being domestic.

Additionally if we were to start over with this project it might have been useful to also analyze more the relationships between the victim and perpetrator in domestic homicides. Doing more analysis on the perpetrator would give us a bigger picture view on domestic homicides because it is not random but rather there is a close relationship between the perpetrator and victim. This would allow deeper analysis on elements such as if fathers are more often the perpetrator. By looking more at this side of the question, our conclusions would contain a broader view of the circumstances relating to domestic homicides, not just that of the victim.