# Analyzing US Murder Data (2015 - 2016)

*Will Arliss & Jackson Turek*

*April 23, 2019*

## Introduction

The data we analyzed is a report of the number of murders from 79 cities in the years 2015 and 2016. The data came from several sources, most of which were police departments (sources and dates can be found in the original data set). The original data set includes the name of the city, name of the state, number of murders in 2015, number of murders in 2016, the change between the two years, a source, and the source date. Additionally, we wanted information on the city populations, so in excel we added a new column for population and entered each city's respective information from 2016, as reported by the Census Bureau. These 8 columns are what were read into RStudio. An important note is that these are only the murders reported to/by the police, and there is potential for murders to have gone unreported. So, although these reported numbers are close, they likely do not capture every single murder in these cities. We can, however, use these numbers as a sample from which we can draw general assumptions and conclusions about murder rates across the country. Some questions we had prior to the start of our analysis included the following: which cities had the highest/lowest quantities of murders per year, which cities had the highest/lowest murder rates per year, which cities had the greatest/least change in murder rates. By analyzing which cities had the highest murder rates year to year, we can begin to focus on those cities and think about possible causes and potential solutions to the crime problems. Cities with extremely low murder rates can be looked to as a model on how to reduce murder rates in problem cities. By looking at the change in murder rates from one year to the next, we can analyze any improvements and failures in policing, legislation, or social behaviors that could account for lower or higher changes in murder rates. Some initial hypotheses we have include: 1) Larger cities will have higher murder rates per year, 2) Chicago, New York, and Los Angeles will have the highest murder rates per year, 3) Change in murder rates from 2015 to 2016 will be the greatest in small cities where policing and other measures to reduce murder have the most impact.

## Tidy Data

```
m1 <- m %>%
  mutate(region = ifelse(state %in% c('Connecticut', 'Maine', 'Massachusetts',
                                      'New Hampshire', 'Rhode Island', 'Vermont',
                                      'New Jersey', 'New York', 'Pennsylvania',
                                      'Delaware', 'Maryland', 'D.C.'),
                         'Northeast', state),
         region = ifelse(state %in% c('Illinois', 'Indiana', 'Michigan', 'Ohio',
                                      'Wisconsin', 'Iowa', 'Kansas', 'Minnesota',
                                      'Missouri', 'Nebraska', 'North Dakota',
                                      'South Dakota'),
                         'Midwest', region),
         region = ifelse(state %in% c('Florida', 'Georgia', 'West Virginia',
                                      'Kentucky', 'North Carolina', 'South Carolina',
                                      'Virginia', 'Tennessee', 'Alabama',
                                      'Mississippi', 'Arkansas', 'Louisiana'),
                         'Southeast', region),
         region = ifelse(state %in% c('Colorado', 'Wyoming', 'Montana', 'Idaho',
                                      'Washington', 'Oregon', 'Utah', 'Nevada',
                                      'California', 'Alaska', 'Hawaii'),
                         'West', region),
         region = ifelse(state %in% c('Texas', 'Oklahoma', 'New Mexico', 'Arizona'),
                         'Southwest', region)) %>%
  gather(year, murders, X2015_murders, X2016_murders)
```

The original data set, read in as "m", was not tidy. It contained two columns for murders by year when it could have had just one column for murders and another for year. The gather() verb in tidyr was used to fix this. Additionally, we wanted to group each city by region in order to perform broader analyses. We did this using the mutate() verb in dplyr five times to classify states into each region. Note that "m1" is not the data set used in our analysis. "m1" was used to create columns that can be put into the final data set.

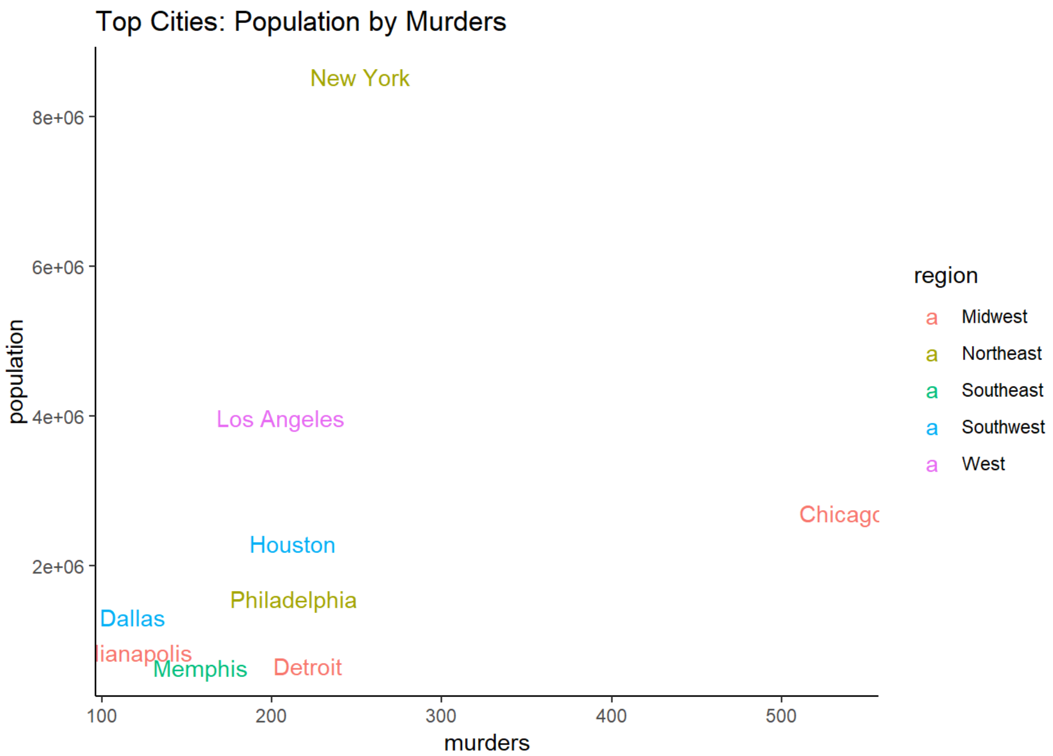
```
MurderData <- m %>%
  mutate(murder_rate_15 = (X2015_murders*10000) / population,
         murder_rate_16 = (X2016_murders*10000) / population,
         murder_rate_change = (murder_rate_16 - murder_rate_15)) %>%
  gather(year, murder_rate, murder_rate_15, murder_rate_16) %>%
  mutate(year = ifelse(year == 'murder_rate_15', '2015', '2016'),
         murders = m1$murders, region = m1$region) %>%
  select(year, city, state, region, population, murders,
         murder_rate, change, murder_rate_change)
```

Here, the final data set, "MurderData", is also read from the original data set, "m". This is because we had to use "key = year" in the gather() function two times in order to get a column for murders per year and murder_rate by year (doing so in the same data set would have created unnecessary duplicates). Additionally, in order to calculate murder rates, we needed the original "X2015_murders" and "X2016_murders" columns. Murder rates for both years are calculated by multiplying the number of murders by 10,000 then dividing it by population. This provides a statistic that can be read as: "There were x-amount of murders for every 10,000 people in this city." Subtracting those two values gave us the "murder_rate_change" column. After gathering the year and "murder_rate" columns, the values in the "year" column had to be changed from "murder_rate_x" to an actual year. This was done using an ifelse statement in the mutate() function. The years were entered as strings to make graphing more clear. Next, the "region"" and number of murders ("murders") columns were pulled into the "MurderData" set from the "m1" data set. Finally, we narrowed down our data set to only display "year", "city", "state", "region", "population", "murders", "murder_rate", "change", and "murder_rate_change".

## Data Analysis

### Graph 1: Top Cities: Population by Murders

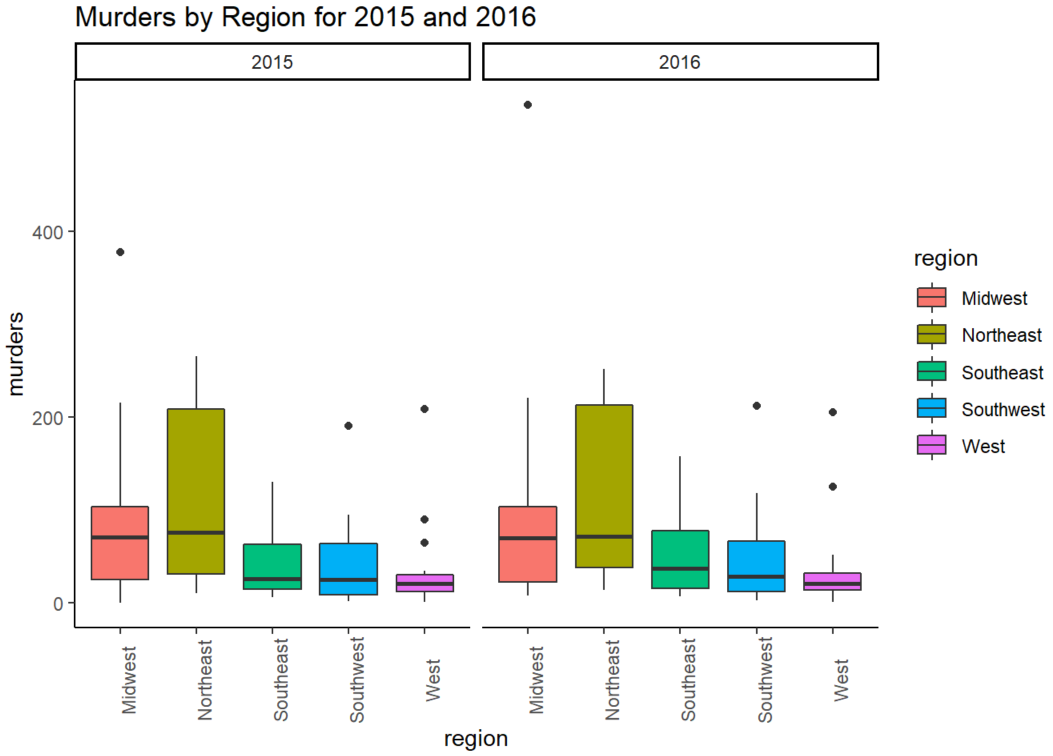
```r
MurderData %>%
  filter(year == '2016',
         murders > 115,
         population > 650000) %>%
  ggplot(mapping = aes(x = murders,
                       y = population,
                       color = region)) +
  geom_text(aes(label = city)) +
  theme_classic() +
  ggtitle("Top Cities: Population by Murders")
```




This graph shows a relationship between population size and the quantity of murders associated with that population in different cities. Here, we can see that although New York has a higher population than Chicago, Chicago has more murders. This allows us to conclude that there is not a linear relationship between population and murder rates. The implications of this graph are that cities with higher populations tend to have higher quantities of murder. However, one city having a high population does not automatically mean that it will have more murders than another city. This makes sense because if there are more people to commit murder and get murdered, there will be higher quantities of murder. This is why we were surpised to find that Chicago had more murders than New York, despite having a smaller population.

## Graph 2: Box Plot of Murders by Region for 2015 and 2016

```r
ggplot(data = MurderData,
       mapping = aes(x = region,
                     y = murders,
                     fill = region)) +
  geom_boxplot() + theme_classic() +
  facet_wrap(~year, ncol = 2) +
  theme(axis.text.x = element_text(angle = 90)) +
  ggtitle("Murders by Region for 2015 and 2016")
```




This graph shows a box plot of murders by region (of observed cities) for 2015 and 2016. In 2015, we can see that Northeastern cities have the largest range of murder quantities and Western cities have the smallest range. The size of each region's range demonstrates the variability in the quantity of murders per city. This graph is interesting, as it displays consistency (or lack there of) of murder quantities in each region. This begs the question: why would cities in one region vary more in murder quantities than cities in a different region? The answer to this likely requires information on policing and public legislatiure. Furthermore, in 2016 we see that Northeastern cities still have the largest range of murder quantities, and Western cities still have the smallest range. What is interesting in this comparison is that in 2016, the Northeast was the only region to report a (slightly) lower mean than in 2015. The implications of this graph are that either the sizes of cities in the Northeast vary greatly, or the legistlative/policing methods in the Northeast cities vary. The fact that the mean of murders seen in Northeast cities went down implies that there were fewer murders for some reason. This could be due to a decrease in population or more effective policing/legislative policies that were sucessful in reducing murders. It should also be noted that the data set could contain more statistics from the Northeast than from other regions.

## Top 10 Cities with Increase in Murder Rate per 10,000 People

```
MurderData %>%
  filter(year == 2016) %>%
  arrange(desc(murder_rate_change)) %>%
  select(city, population, change, murder_rate_change) %>%
  head(10)
```

| |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

1-10 of 10 rows | 1-1 of 5 columns

This table shows 10 cities with the greatest increases in murder rates per 10,000 people. The table allows us to empirically analyze the largest positive changes in murder rates from 2015 to 2016 across cities, using the same sample size. A positive change in murder rate means that a city's 2016 murder rates were higher than those in 2015. Orlando and Jacksonville had the most significant increases in murder rates. The difference between the second city and the third city (Memphis) was the greatest difference seen in the entire table. By using a regulated population of 10,000 we were able to compare the statistics across all cities, which gave a more accurate representation of the increase in crime (possibly due to the failure of policing/legislature).

## Top 10 Cities with Decrease in Murder Rate per 10,000 People

```
MurderData %>%
  filter(year == 2016) %>%
  arrange(murder_rate_change) %>%
  select(city, population, change, murder_rate_change) %>%
  head(10)
```

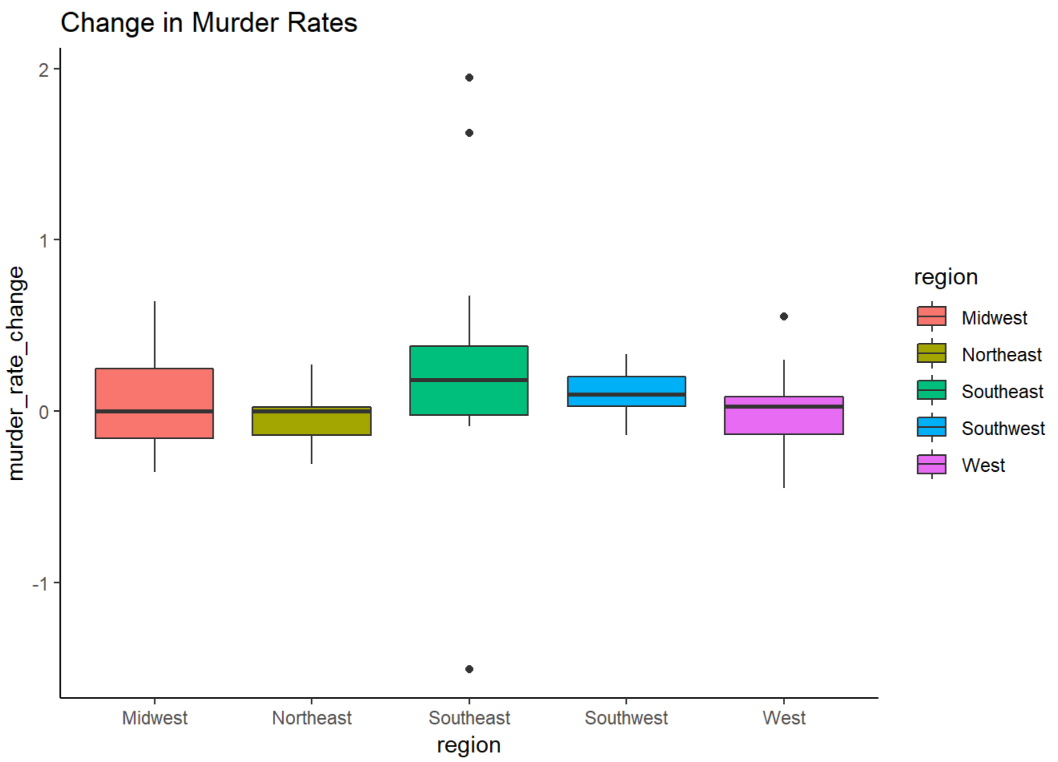
| |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

1-10 of 10 rows | 1-1 of 5 columns

This table shows 10 cities with the greatest decreases in murder rates per 10,000 people. The table allows us to empirically analyze the largest negative changes in murder rates from 2015 to 2016 across cities, using the same sample size. Miami showed by far the largest decrease in murder rate, as seen by the large difference in murder rates from Miami to the second city (San Francisco). By using a regulated population of 10,000 we were able to compare the statistics across all cities, which gave a more accurate representation of the decrease in crime (possibly due sucess in policing/legislature).

### Graph 3: Box Plot of Change in Murder Rate

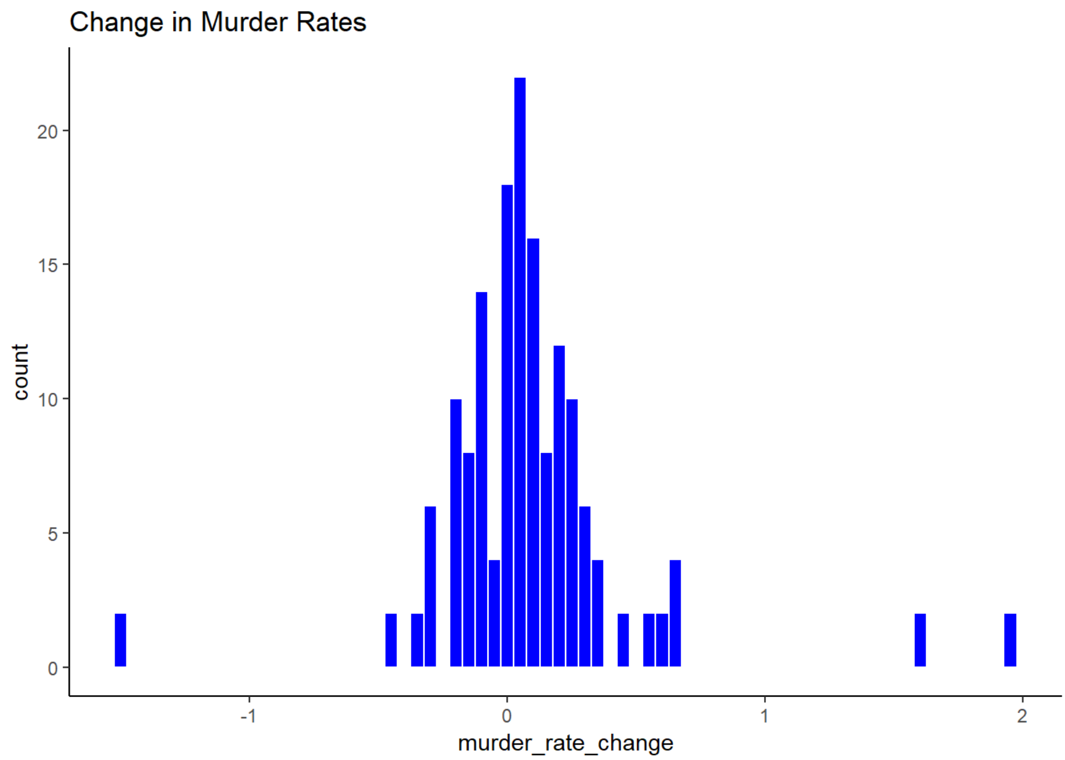
```
ggplot(data = MurderData,
       mapping = aes(x = region,
                     y = murder_rate_change,
                     fill = region)) +
  geom_boxplot() + theme_classic() +
  ggtitle("Change in Murder Rates")
```




This graph shows the change in murder rate by region. The mean for Northeast went down from 2015 to 2016, and every other region went up. The Southeast seems to have gone up the most, indicating that they did the poorest job in combatting murder rates over the course of the year. This graph allows us to visually analyze changes in murder rates per 10,000 people across all of the cities in our data set (grouped by region), thus implicating which region contained cities that were the most or least successful in combatting murder from 2015 to 2016.

## Graph 4: Histogram of Change in Murder Rate

```
ggplot(data = MurderData,
       mapping = aes(x = murder_rate_change)) +
  geom_histogram(fill = 'blue',
                 color = 'white',
                 bins = 70) +
  theme_classic() +
  ggtitle("Change in Murder Rates")
```




This graph looks at the number of cities across the entire country that were contained within certain ranges of change in murder rate. This graph is precise and centered around zero. It is unimodal and symmetric. The implications of this graph are that most cities are inclined to have a low change in murder rate. This means that most cities are not going to see dramatic change in murder rate and that most cities will generally see an increase in murders from one year to the next. This makes sense becuase a one year period is a very short time to introduce policy that would provide significant changes, and it is unlikely anything would change significantly by chance.

## Top 10 Cities by Murder Rates in 2015 and 2016

```
HighestRates15 <- MurderData %>%
  filter(year == '2015') %>%
  arrange(desc(murder_rate)) %>%
  select(year, city, population, murders, murder_rate) %>%
  head(10)

HighestRates16 <- MurderData %>%
  filter(year == '2016') %>%
  arrange(desc(murder_rate)) %>%
  select(year, city, population, murders, murder_rate) %>%
  head(10)
```

```
HighestRates15
```

| |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

1-10 of 10 rows | 1-1 of 6 columns

```
HighestRates16
```

| |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

1-10 of 10 rows | 1-1 of 6 columns

These tables show 10 cities with the highest murder rates per 10,000 people in 2015 and in 2016. The tables demonstrate how many people in any sample of 10,000 people could be murdered in one year, depending on the city from which the sample is drawn. This is important because it demonstrates the real crime rate and the relative danger of living in any of these cities. The cities contained in the list do not seem to change too dramatically between 2015 and 2016.

## Lowest 10 Cities by Murder Rates in 2015 and 2016

```
LowestRates15 <- MurderData %>%
  filter(year == '2015') %>%
  arrange(murder_rate) %>%
  select(year, city, population, murders, murder_rate) %>%
  head(10)

LowestRates16 <- MurderData %>%
  filter(year == '2016') %>%
  arrange(murder_rate) %>%
  select(year, city, population, murders, murder_rate) %>%
  head(10)
```

```
LowestRates15
```

| |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

1-10 of 10 rows | 1-1 of 6 columns

```
LowestRates16
```

| |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8 |
| 9 |
| 10 |

1-10 of 10 rows | 1-1 of 6 columns

These tables show 10 cities with the lowest murder rates per 10,000 people in 2015 and in 2016. The tables demonstrate how many people in any sample of 10,000 people could be murdered in one year, depending on the city from which the sample is drawn. This is important because it demonstrates the real crime rate and the danger of living in any of these cities. These tables display significantly lower rates than the previous two. The cities contained in the list here seem to vary a bit between 2015 and 2016.

# Conclusion

At the start of our data analysis, we had three hypotheses: 1) Larger cities will have higher murder rates per year, 2) Chicago, New York, and Los Angeles will have the highest murder rates per year, 3) Change in murder rates from 2015 to 2016 will be the greatest in small cities where policing and other measures to reduce murder will have the most impact. We can reject hypothesis 1 because we saw that cities such as Jacksonville and San Francisco topped the charts with high murder rates both years, and they have relatively small populations when considering other cities in our study. We can reject hypothesis 2 because none of these cities were in the top 10 highest murder rates in either year. However, when not considering population, they have among the highest quantities of murder. We can also reject hypothesis 3 because, although Jacksonville is in the top 10 decrease in murder rates, it is the only relatively small city in that group. Overall, we cannot extrapolate much from this data because we do not have information on why murder rates went up or down. All we know is that they did go up or down, which could be for many reasons. It is unsettling to think that these variations could simply be due to complete randomness. The analysis of this data was most sucessfull in looking at the true violence of each city in the sample we were given. By setting a standardized population of 10,000 people, we were able to see the amount of murders in each city per that amount of people. This data can shed light on which cities were actually more dangerous in terms of murders for a constant amount of people.

# Limitations

One key limitation that existed in our data was unreported crime. There are a lot of murders in any city that go unreported, undiscovered, or are not treated as actual murder. The data we used can not be looked at as the true number of murders, but rather an estimate. It would have been interesting to have data on policing/legislative techniques to go with the statistics on each city’s murders. Another limitation is the possibility that some regions are over-represented in the data set than others. It was observed that Northeastern cities have a much larger range in quantity of murders than Western cities. The reason for this could be that there were more Northeastern murder statistics recorded in the data set than Western statistics, and therfore the Northeastern data was more spread out.