Code ▾

DATS610 Midterm Project- Analyzing trends in U.S. Homicide Data

Junran Cao

2021-11-07

1 Background

1.1 Overview

Homicide is defined as the act of a human being killing another human. Our group wanted to study the topic of homicides in the United States for several reasons. While homicides are obviously tragic and unfortunate, the reality is that we are naturally curious about the circumstances surrounding a homicide. We ask questions such as: who was the victim? Who committed the homicide, and what was their motivation for doing so? What type of weather that was? The sheer number of homicide cases in the United States in recent history provides a huge amount of data at our disposal. We can analyze both solved and unsolved homicide cases to answer questions about the victims, the perpetrators, and many other aspects of homicide.

The data that we need in order to thoroughly analyze homicide cases in the United States is available thanks to the Murder Accountability Project, a non-profit dedicated to compiling and tracking unsolved homicides nationwide. Founded by investigative journalist and former White House correspondent Thomas Hargrove, the Murder Accountability Project keeps extensive records on homicides and the relevant information surrounding them by gathering information from the FBI, the Department of Justice, and other agencies. The result of their work is a massive set of data that lists relevant information for over 600,000 homicide cases that occurred between 1980 and 2014. The data set includes information on both the victim and perpetrator (sex, age, race, ethnicity) as well as the department that handled the case. We have access to the relationship between the victim and perpetrator, the weapon used in the homicide, and whether that homicide was solved. In addition to the homicide data, we also have access to monthly weather temperature data courtesy of the National Oceanic and Atmospheric Administration that we can use to look at the relationship, if any, between weather and homicide cases.

1.2 Prior Researches

We found some prior researches on this data set. They were mostly researching on the demographic information of the victim and perpetrator. There were some researches on demographic information and the weapon type or type of homicide. Also, some researches based on yearly homicide cases, predicting demographic of victim/perpetrator.

1.3 Dataset Description

Serial No.	Column Name	data type	description
1	Record.ID	Integer	Unique ID of Record
2	Agency.Code	Integer	Code of the agency handling the case
3	Agency.Name	Integer	Name of the agency handling the case
4	Agency.Type	Character	Type of the agency handling the case
5	City	Character	City name of the homicide incident
6	State	Character	State of the homicide incident
7	Year	Integer	Year of the homicide incident
8	Month	Character	Month of the homicide incident
9	Incident	Integer	Code of the Incident type
10	Crime.Type	Character	Type of crime
11	Crime.Solved	Character	Crime solved or not
12	Victim.Sex	Character	Sex of Victim
13	Victim.Age	Integer	Age of Victim
14	Victim.Race	Character	Race of victim
15	Victim.Ethnicity	Character	Victim Ethnic group
16	Perpetrator.Sex	Character	Sex of perpetrator
17	Perpetrator.Age	Integer	Age of perpetrator
18	Perpetrator.Race	Character	Perpetrator racial group
19	Perpetrator.Ethnicity	Character	Perpetrator Ethnic group
20	Relationship	Character	Relationship between victim and perpetrator
21	Weapon	Character	Type of weapon used
22	Victim.Count	Integer	count of victim

Serial No.	Column Name	data type	description
23	Perpetrator.Count	Integer	Count of perpetrator
24	Record.Source	Character	Source of the homicide report
25	temp	Numeric	Monthly average temperature
26	Region	Character	Region of states

1.4 Limitation of the dataset

In the data set, there isn’t any specific record of the date of the incident. It would have been better if there was exact incident date information. While going through the scenario of homicides, we decided to find out if there could be any relationship between homicides and weather temperature. To achieve that we needed the temperature data of the homicide cases. We found the monthly average data from the NOAA website. We found three data files over there, Hawaii monthly average temperature data, Alaska monthly average temperature data and another with all other 48 states monthly average temperature data. Joining these three separate data sets, We made a data file with all the 50 states monthly average data and finally joined the monthly average column as ‘temp’ to our main homicide data set. If we had the information of the exact homicide incident date, we could use that day’s average temperature instead of the monthly average temperature. Then the study would have been more accurately compared.

A region column has been added so that the data can be analyzed on the regional level in addition to the state level. The data is based on the regions of the U.S. as defined by the U.S. Census Bureau (U.S. Census Bureau).

1.5 SMART Questions and Hypotheses

After finding an accessible source of data for homicide cases, there are several relevant questions we can ask about the nature and circumstances surrounding homicides.

1. Is there a relationship between the weapon used and relationship with the victim?

Since we have access to the victim/perpetrator relationship as well as the weapon used in the homicide, we are curious to explore if these two variables are related. Our hypothesis is that the two are related based on how close the victim and perpetrator were to each other. For example, if the victim and perpetrator were romantic partners or close family members, the homicide was more likely to have been committed using an intimate weapon, such as a blunt melee object, drowning, or strangulation. On the other hand, if the victim and perpetrator were not close with each other, the homicide was more likely to involve a ranged weapon such as a firearm or explosive. We are curious to see if the homicide data supports or refutes this hypothesis.

2. What is the relationship, if any, between perpetrator and victim?

Explanation for question 2.

3. Which regions of the USA experience higher rates of homicide?

Our data set includes which state each homicide occurred in so we are going to explore if there is a region of the country that has more homicides than the others. Along with this question we are going to explore the possible relationship between the region of the country and how many victims were involved in each homicide.

4. Is there a relationship between the year of the homicide and whether or not the case has been solved?

We have data that spans from 1980 to 2014 so we would like to find out if there are any patterns behind how many cases get solved. Our hypothesis is that as we progress through the years the proportion of solved to unsolved cases will grow. Technology has been advancing very quickly the last few decades so we want to find out if the amount of cases solved increases as technology advances.

5. Which states are best at solving homicide cases?

To find out which state is best at solving case at first we want to find out which are the states that have more total homicide cases. As those states are having more number of cases, it will be more informative to find out how much efficient they are at solving cases. We intend to find out the success rates of the top 5 states with most homicide cases.

6. Is there a relationship between weather type/temperature and number of homicides committed?

There had been several researches going on since decades to find out the correlation between weather and crimes. There is a strong hypothesis that they have a great interrelationship among them. Well, of course, the weather doesn't explicitly cause crime. It is people's actions that lead to violence. But we intent to find out if our data set can show some connection between the number of homicides and the weather temperature. Can there be any correlation? To match up the original data set, we are using monthly average temperature to analyze.

After establishing some relevant questions pertaining to the data set, we are now ready to explore the data itself in order to gain insight into these questions.

2 Analysis

2.1 Exploratory Data Analysis (EDA)

2.1.1 Weapon and Type of relationship between victim and perpetrator

Hide

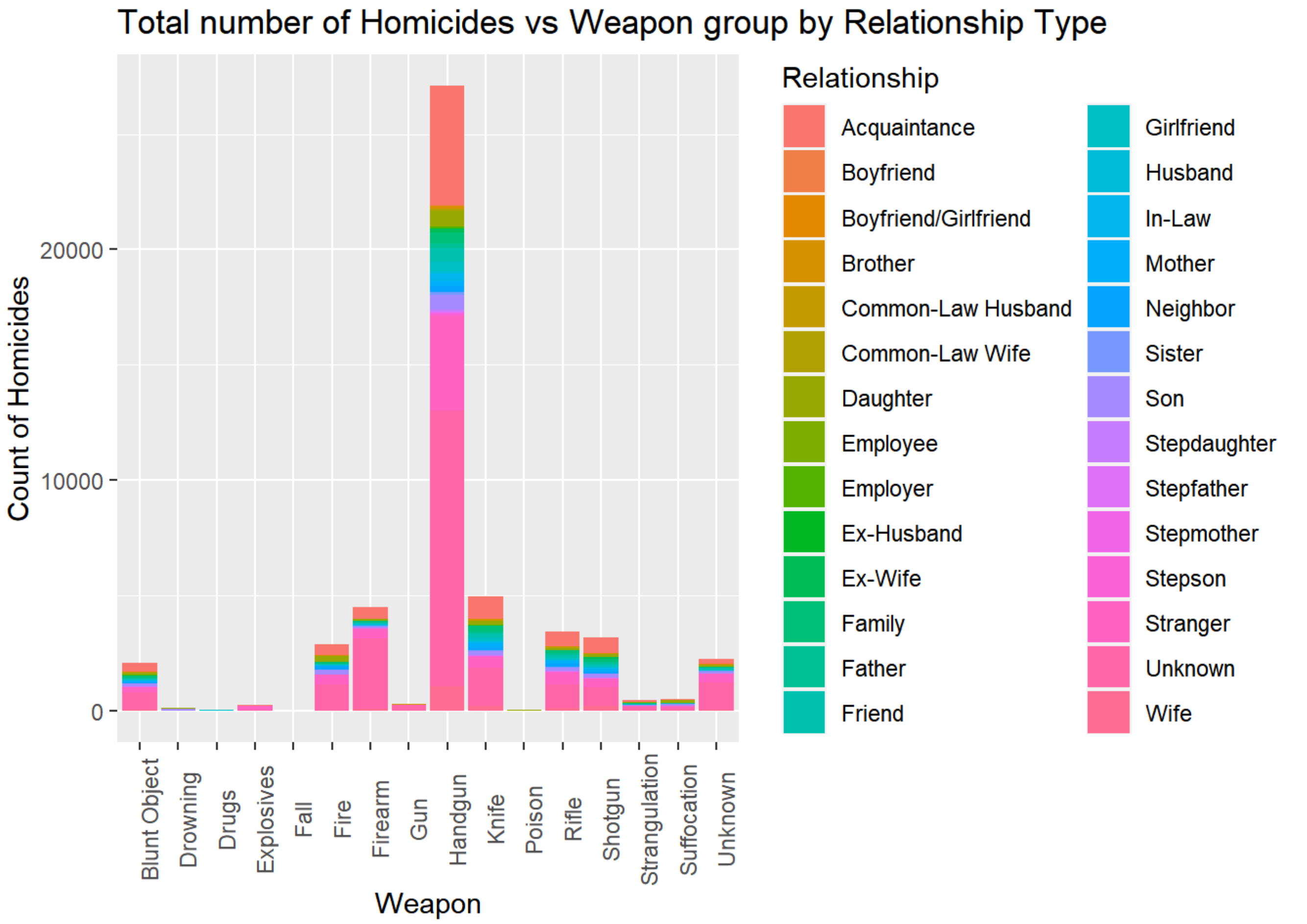
```
crime <- data.frame(read.csv(unz('homicide_data/homicide_data.zip','homicide_data.csv'), header = T))
head(crime, 5)
```

The dataframe created by combining the homicide data and the monthly weather data from NOAA contains 27 variables and 638454 observations.

In order to explore the relationship between the weapon used in the homicide and the victim/perpetrator relationship, we will first filter the data set to include only the cases with at least one victim.

Hide

```
#EDA
homicides = crime %>% filter(Victim.Count > 0)
ggplot(homicides, aes(x = Weapon, fill = Relationship)) +
  geom_bar(position = "stack") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title="Total number of Homicides vs Weapon group by Relationship Type", y="Count of Homicides")
```



As we can see from the graph above, there are simply too many types of relationships between the victim and perpetrator to properly analyze the interactions between variables. In order to streamline our analysis, we chose to combine the different relationships into 5 different factors: Romantic Partner, Family Member, Friend/Acquaintance, Coworker, and Stranger/Unknown. We will fit the 28 existing types of relationships into the following categories and visualize the data in a more effective manner.

Coworker: Employer, Employee

Family Member: Brother/Sister, Son/Daughter, Family, Father/Mother, In-law, Stepson/Daughter, Stepmother/father.

Friend/Acquaintance: Friend, Acquaintance, Neighbor

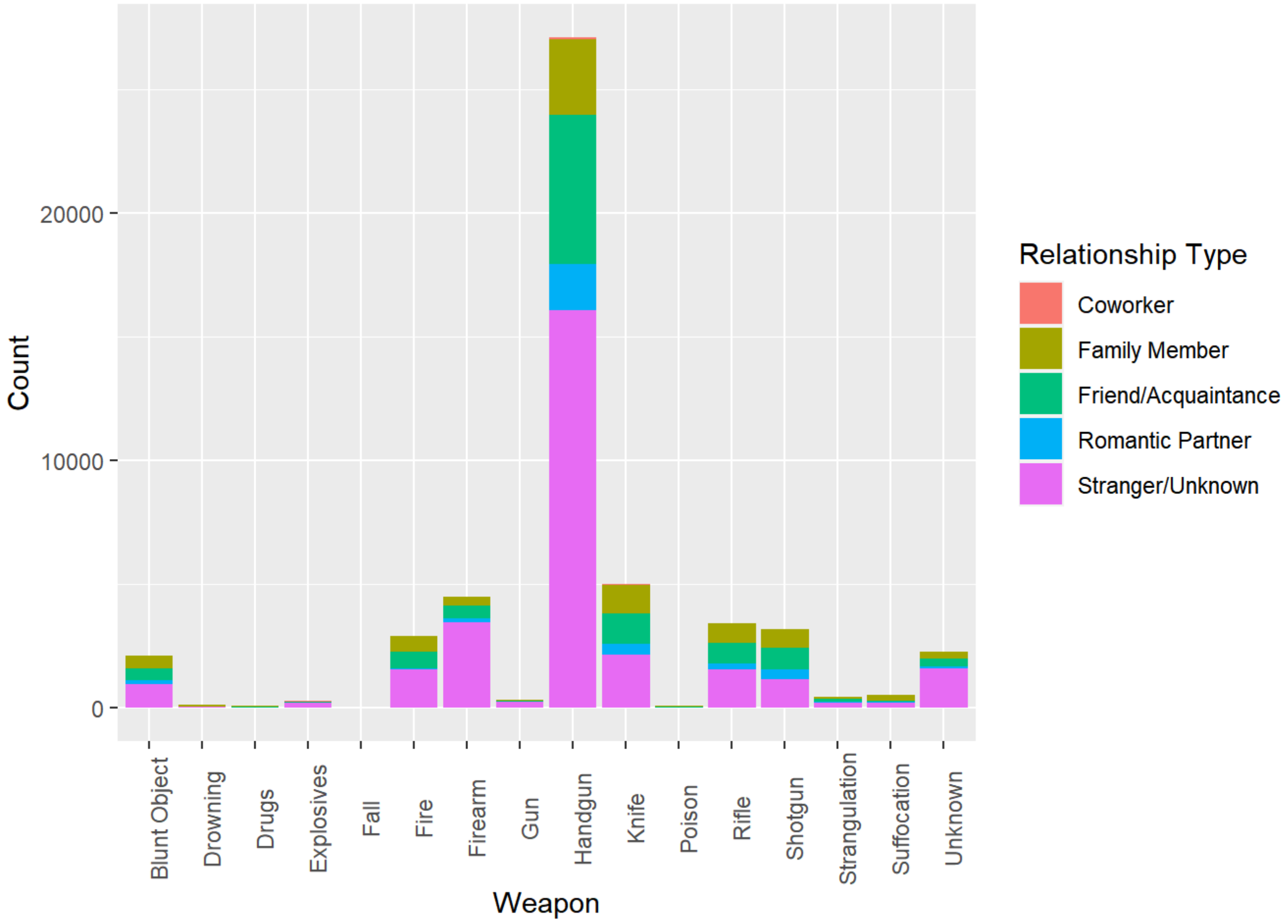
Romantic Partner: Boyfriend, Girlfriend, Boyfriend/Girlfriend, Common-Law Husband/Wife, Ex-Husband/Wife, Husband/Wife.

Stranger/Unknown: Stranger and Unknown

Hide

```
#creating relationship type column
homicides$Relationship_Type[homicides$Relationship == 'Boyfriend'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Girlfriend'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Boyfriend/Girlfriend'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Common-Law Husband'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Common-Law Wife'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Ex-Husband'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Ex-Wife'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Husband'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Wife'] <- 'Romantic Partner'
homicides$Relationship_Type[homicides$Relationship == 'Stranger'] <- 'Stranger/Unknown'
homicides$Relationship_Type[homicides$Relationship == 'Unknown'] <- 'Stranger/Unknown'
homicides$Relationship_Type[homicides$Relationship == 'Brother'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Daughter'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Family'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Father'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'In-Law'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Mother'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Sister'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Son'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Stepdaughter'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Stepfather'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Stepmother'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Stepson'] <- 'Family Member'
homicides$Relationship_Type[homicides$Relationship == 'Employee'] <- 'Coworker'
homicides$Relationship_Type[homicides$Relationship == 'Employer'] <- 'Coworker'
homicides$Relationship_Type[homicides$Relationship == 'Neighbor'] <- 'Friend/Acquaintance'
homicides$Relationship_Type[homicides$Relationship == 'Acquaintance'] <- 'Friend/Acquaintance'
homicides$Relationship_Type[homicides$Relationship == 'Friend'] <- 'Friend/Acquaintance'

#relationship type vs weapon used EDA
ggplot(homicides, aes(x = Weapon, fill = Relationship_Type)) + geom_bar(position = "stack") + theme(axis.text.x = element_text(angle = 90)) + labs(x = 'Weapon', y = 'Count', fill = 'Relationship Type')
```



We can’t draw any definitive conclusions from the graph alone, but it is notable that the vast majority of homicides are committed via handgun.

2.1.2 The relationship, if any, between perpetrator and victim

2.2 Cleaning the dataset

Hide

```
crime$Victim.Sex = factor(crime$Victim.Sex)
crime$Victim.Race = factor(crime$Victim.Race)
crime$Perpetrator.Sex=factor(crime$Perpetrator.Sex)
crime$Perpetrator.Race=factor(crime$Perpetrator.Race)

crime$Weapon=factor(crime$Weapon)
crime$Perpetrator.Ethnicity=factor(crime$Perpetrator.Ethnicity)
crime$Victim.Ethnicity=factor(crime$Victim.Ethnicity)
crime$Crime.Type=factor(crime$Crime.Type)
```

Hide

```
crimeclean=na.omit(crime)
crimeclean
homicidesdf = subset(crimeclean,crimeclean$Victim.Count > 0)
#homicides=as.data.frame(homicides)
#cbind(
  #lapply(
    #lapply(homicides, is.na)
    #, sum)
  #)
```

After removing unwanted variables and making sure the variables types are suitable,let’s explore the relationship between two quantitative variables, Victim.Age and Perpetrator.Age. First, let’s see if they are normally distributed.

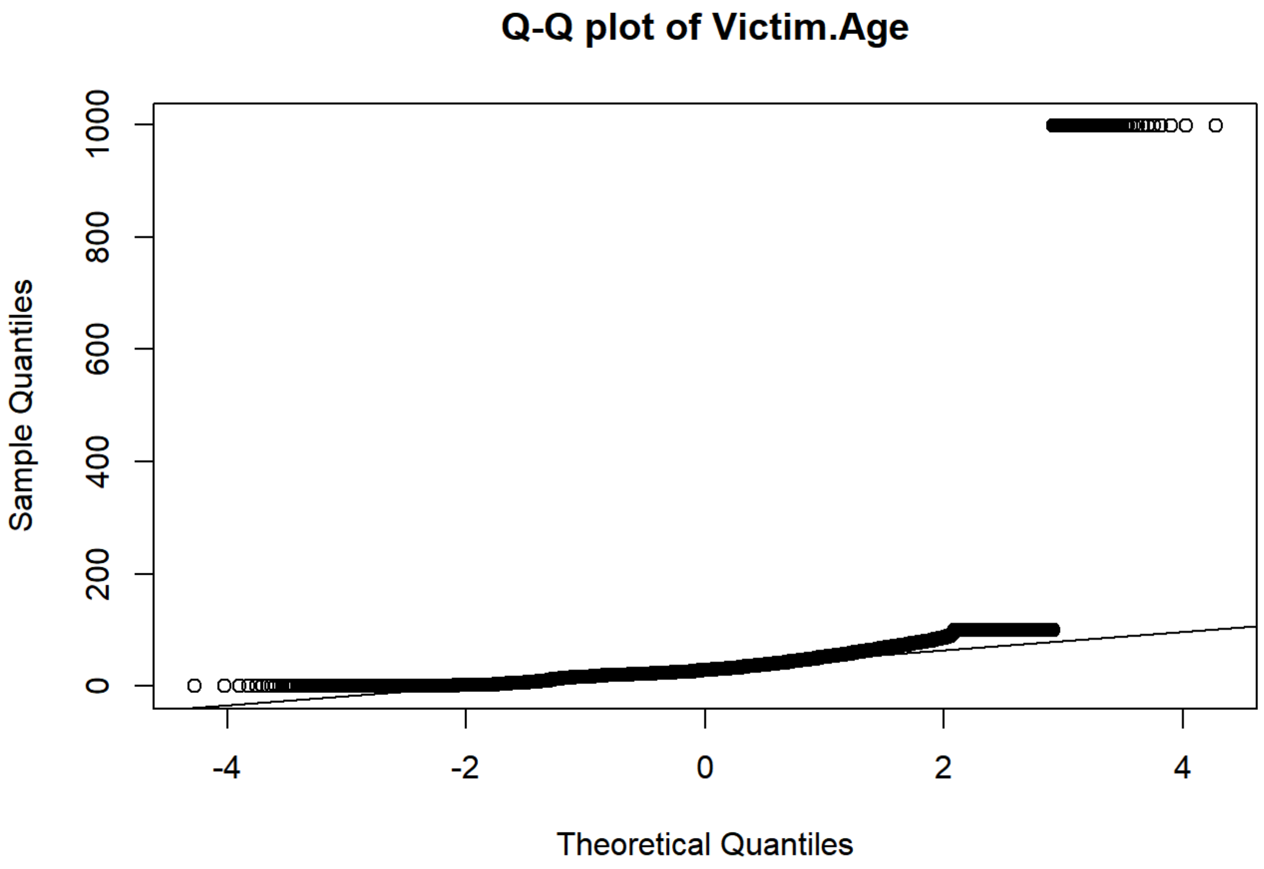
2.3 Normality Check

Hide

```
library(ezids)
```

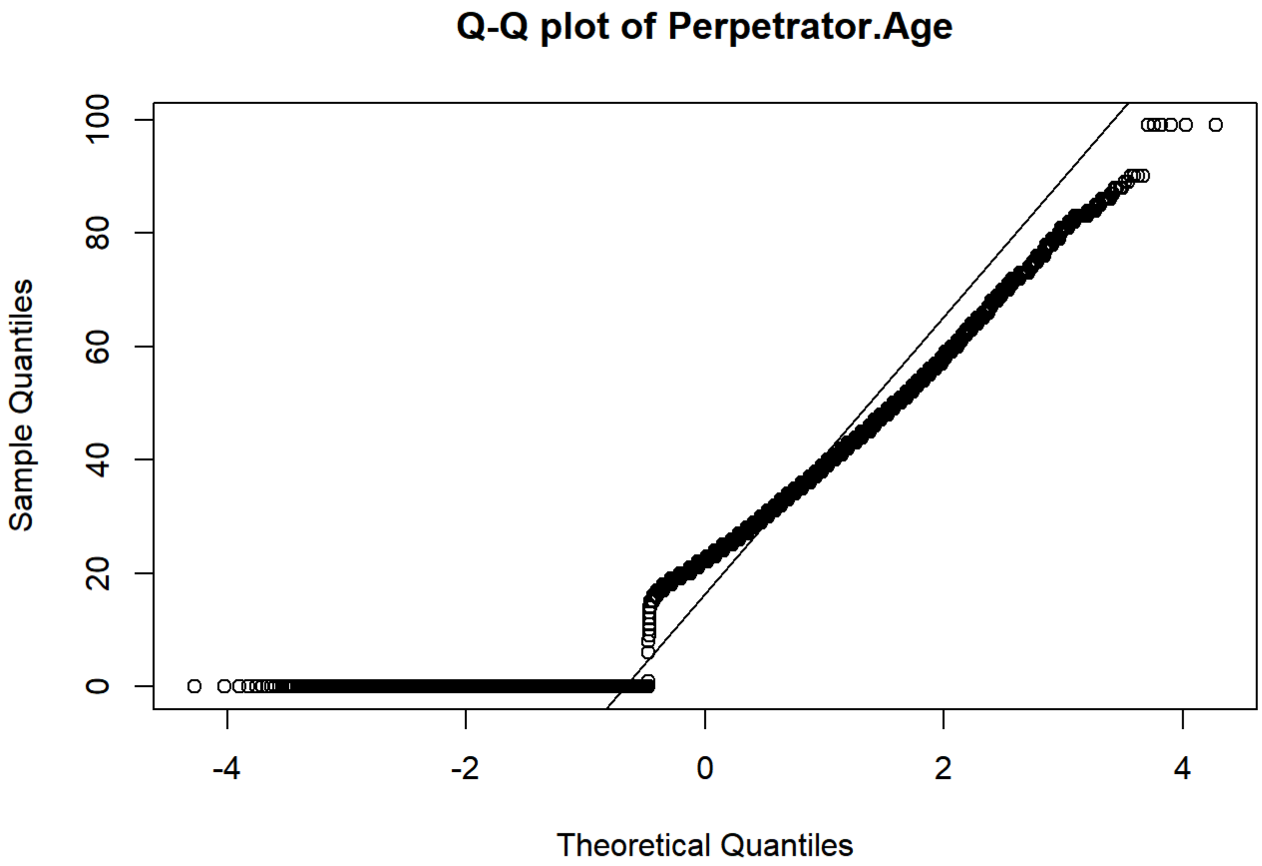
Hide

```
qqnorm(homicidesdf$Victim.Age,main="Q-Q plot of Victim.Age")
qqline(homicidesdf$Victim.Age)
```



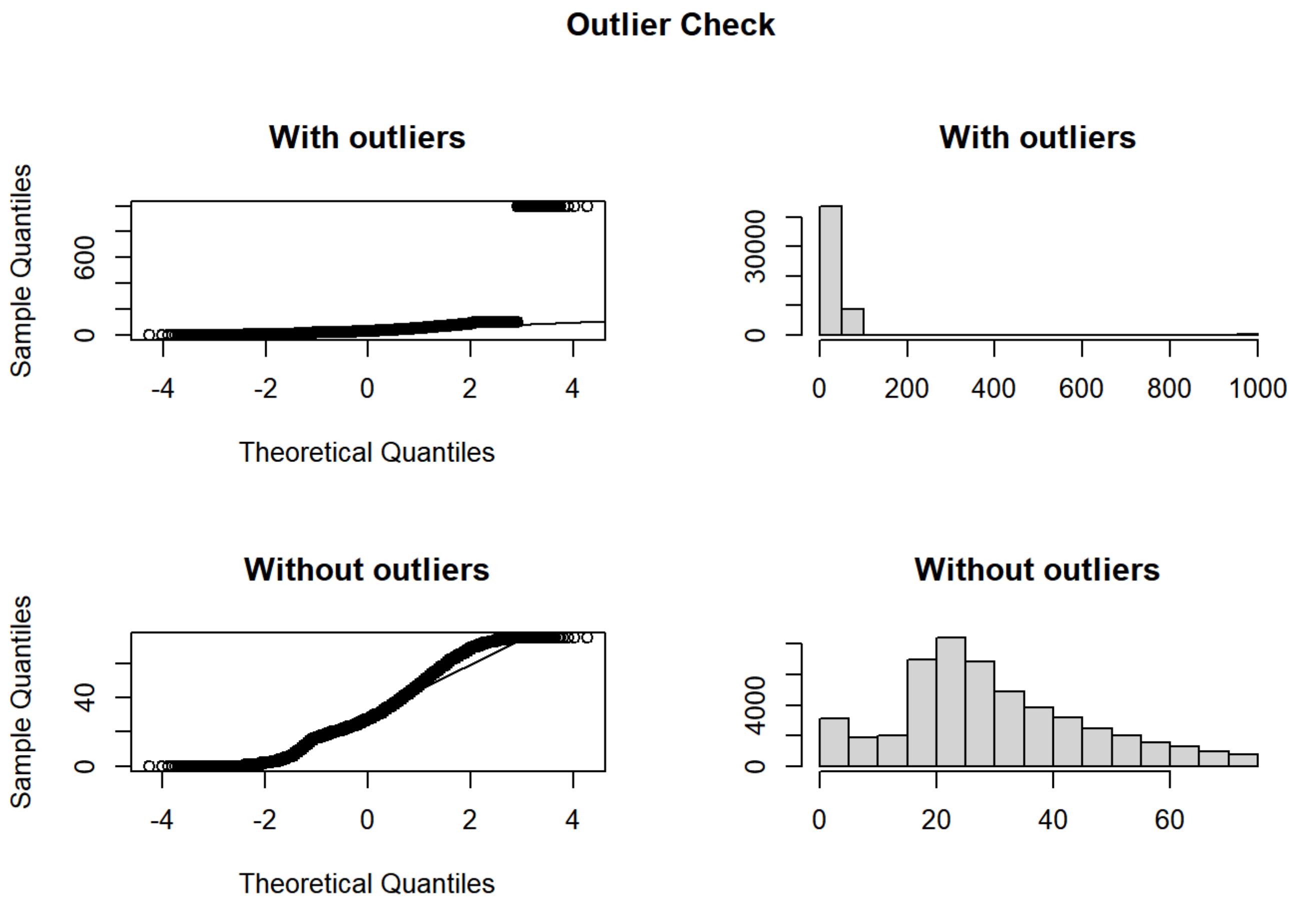
Hide

```
qqnorm(homicidesdf$Perpetrator.Age, main="Q-Q plot of Perpetrator.Age")
qqline(homicidesdf$Perpetrator.Age)
```



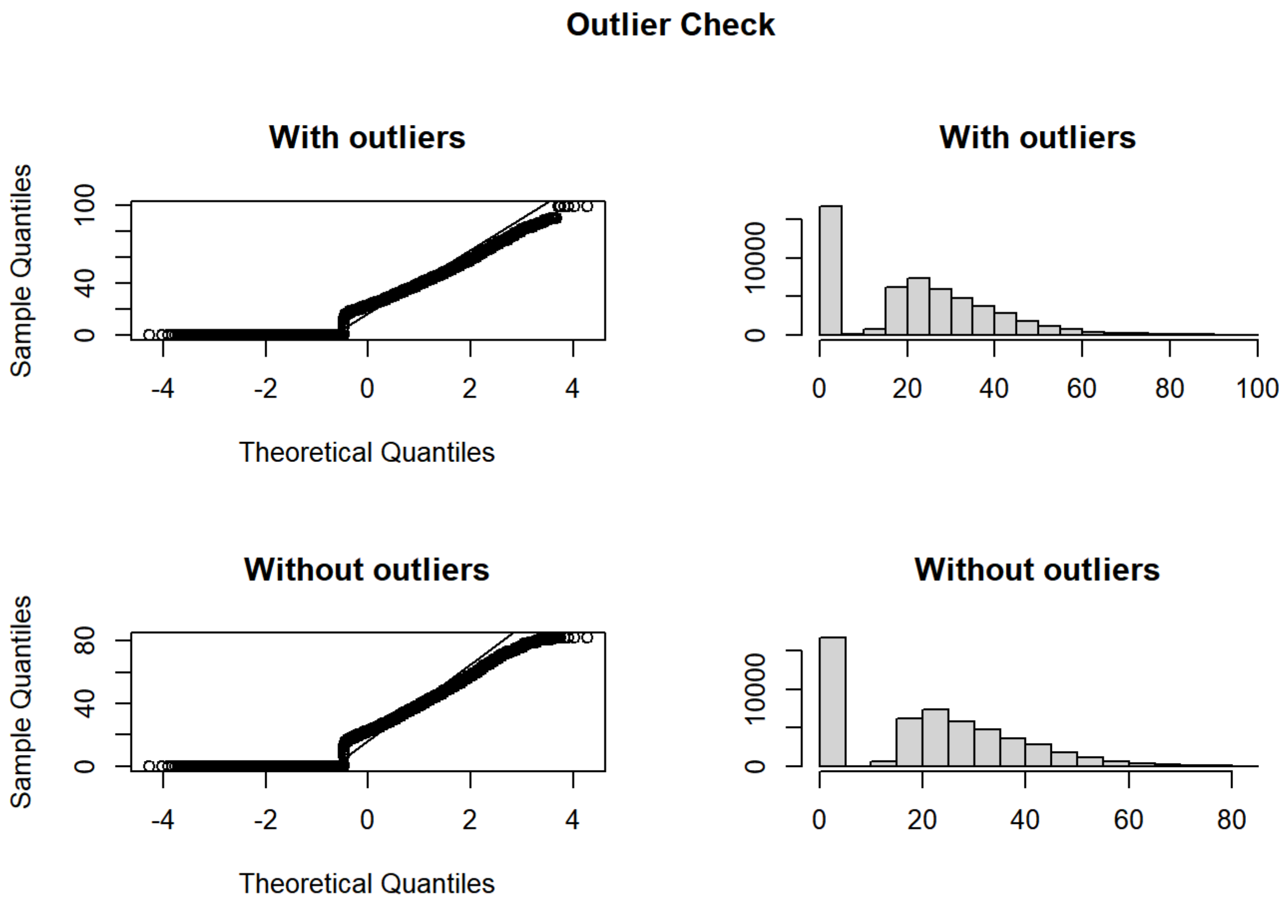
Hide

```
CleanVictimAge = outlierKD2(homicidesdf,homicidesdf$Victim.Age,rm=TRUE,qqplt=TRUE)
```



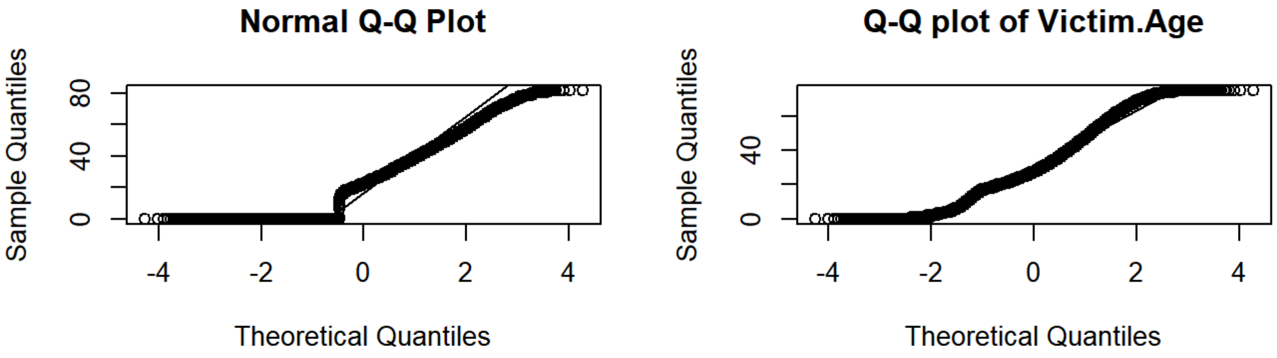
Hide

```
CleanPerpetratorAge = outlierKD2(homicidesdf,homicidesdf$Perpetrator.Age, rm=TRUE,qqplt=TRUE)
```



Hide

```
qqnorm(CleanPerpetratorAge$Perpetrator.Age)
qqline(CleanPerpetratorAge$Perpetrator.Age,main="Q-Q plot of Perpetrator.Age")
qqnorm(CleanVictimAge$Victim.Age, main="Q-Q plot of Victim.Age")
qqline(homicidesdf$Victim.Age)
```


Normal Q-Q Plot

Q-Q plot of Victim.Age

As we can see, after removing the outliers, the qq-plot for Victim.Age improved a lot, meaning we are closer to a normal distribution.However, even without the outliers, the qq-plot for Perpetrator.Age did not improve that much. Since we don't have enough evidence to assume that Perpetrator.Age has a normal distribution, let's test the correlation between the two variables using Spearman's correlation.

2.4 Analysis

Hide

```
cor.test(as.numeric(homicidesdf$Victim.Age),as.numeric(homicidesdf$Perpetrator.Age), method="spearman")
```

Hypothesis Statement: Null hypothesis: Victim age and perpetrator age are not correlated. Alternative hypothesis: Victim age and perpetrator age are correlated

Output: P-value < 0.05

Decision: Reject the null hypothesis and accept the alternative hypothesis

Conclusion: Since the P-value is less than 0.05, we can conclude that victim age and perpetrator age are positively correlated, meaning that victim age is related to perpetrator age in a positive way (rho=0.08258801)

** Let's now look at the relationship between categorical variables

Hide

```
contable = table(homicidesdf$Victim.Race, homicidesdf$Perpetrator.Race)
chitest = chisq.test(contable,simulate.p.value = TRUE)
chitest$statistic
chitest$p.value
chitest$parameter
xkabledply(chitest$expected, title = "Cross table for the expected frequencies between Victim.Race vs Perpetrator.Race")
```

Hypothesis Statement: Null hypothesis: Victim race and perpetrator race are not associated. Alternative hypothesis: Victim race and perpetrator race are associated

Output: P-value < 0.05

Decision: Reject the null hypothesis and accept the alternative hypothesis

Conclusion: With the P-value being less than 0.05, We can reject the null hypothesis and accept the alternative hypothesis. We can therefore conclude that there is a statistically significant association between victim race and perpetrator race. Therefore, we can say that the levels of victim race and perpetrator race vary together, at least a little

Hide

```
contable = table(homicidesdf$Victim.Sex, homicidesdf$Perpetrator.Sex)
chisq.test(contable, simulate.p.value = TRUE)
chitest$statistic
chitest$p.value
chitest$parameter
xkabledply(chitest$expected, title="Cross table for the expected frequencies between Victim.Sex vs Perpetrator.Sex")
```

Hypothesis Statement: Null hypothesis: Victim sex and perpetrator sex are not associated. Alternative hypothesis: Victim sex and perpetrator sex are associated

Output: P-value < 0.05

Decision: Reject the null hypothesis and accept the alternative hypothesis

Conclusion: With the P-value being less than 0.05, We can reject the null hypothesis and accept the alternative hypothesis. We can therefore conclude that there is a statistically significant association between victim sex and perpetrator sex. This result tells us that the levels of victim sex and perpetrator sex vary together, at least a little

Hide

```
contable = table(homicidesdf$Weapon, homicidesdf$Perpetrator.Race)
chisq.test(contable,  simulate.p.value = TRUE)
chitest$statistic
chitest$p.value
chitest$parameter
xkabledply(chitest$expected, title = "Cross table for the expected frequencies between Weapon vs Perpetrator.Race")
```

Hypothesis Statement: Null hypothesis: Weapon used and perpetrator race are not associated. Alternative hypothesis: Weapon used and perpetrator race are associated

Output: P-value < 0.05

Decision: Reject the null hypothesis and accept the alternative hypothesis

Conclusion: With the P-value being less than 0.05, We can reject the null hypothesis and accept the alternative hypothesis. We can therefore conclude that there is a statistically significant association between weapon used and perpetrator race, meaning that the levels of weapons used and perpetrator race vary together, at least a little

** Let’s then try to build some linear models to have some predictor variables predicting the outcome variable (Victim.Age)

Hide

```
mixfit1<- lm(Victim.Age ~ (Perpetrator.Age*Perpetrator.Race), data = homicidesdf)
xkabledply(mixfit1, title = paste("Model:", format(formula(mixfit1)) ) )
vif_md1 = faraway::vif(mixfit1)
vif_md1
xkabledply(mixfit1, title="Summary of LM")
xkablevif(mixfit1, title="VIFs of the model")
```

Looking at the results, we can’t conclude that the slope of the relationship between victim age and perpetrator age to be different depending on perpetrator race (most of the P-values being greater than 0.05). We can see that the correlation between perpetrator age and perpetrator race for both White and Black are very high (their VIF values being greater than 10). Ovoerall, this linear model doesn’t assist our analysis that much given the high level of multicollinearity and the fact that a great portion of the data contains unknown information.

** Let’s add more predictor variables and see how they may predict Victim.Age

Hide

```
mixfit2<- lm(Victim.Age ~ (Perpetrator.Age*Perpetrator.Race+Weapon), data = homicidesdf)
xkabledply(mixfit2, title = paste("Model:", format(formula(mixfit1)) ) )
vif_md2 = faraway::vif(mixfit2)
vif_md2
xkabledply(mixfit2, title="Summary of LM")
xkablevif(mixfit2, title="VIFs of the model")
```

Although this model is exposed to the issues of having high VIF values for Perpetrator.Age and Perpetrator.Race and unknown data, however,we can conclude that drowning, drug, fire, firearm, shotgun, strangulation, and suffocation do contribute significantly to the variance in victim age (their P-values being less than 0.05), which reveals the relationship between the weapon used and the victim age.

2.5 Main Findings

With a great portion of the data containing unknown information, the relationships between victim and perpetrator are comparatively vague in general. However, we can conclude that:

- The victim and perpetrator are likely from the same racial and/or age group
- Certain weapons contribute significantly to the variance in victim age, a relationship between the vitality of the weapons and the victim age is seen

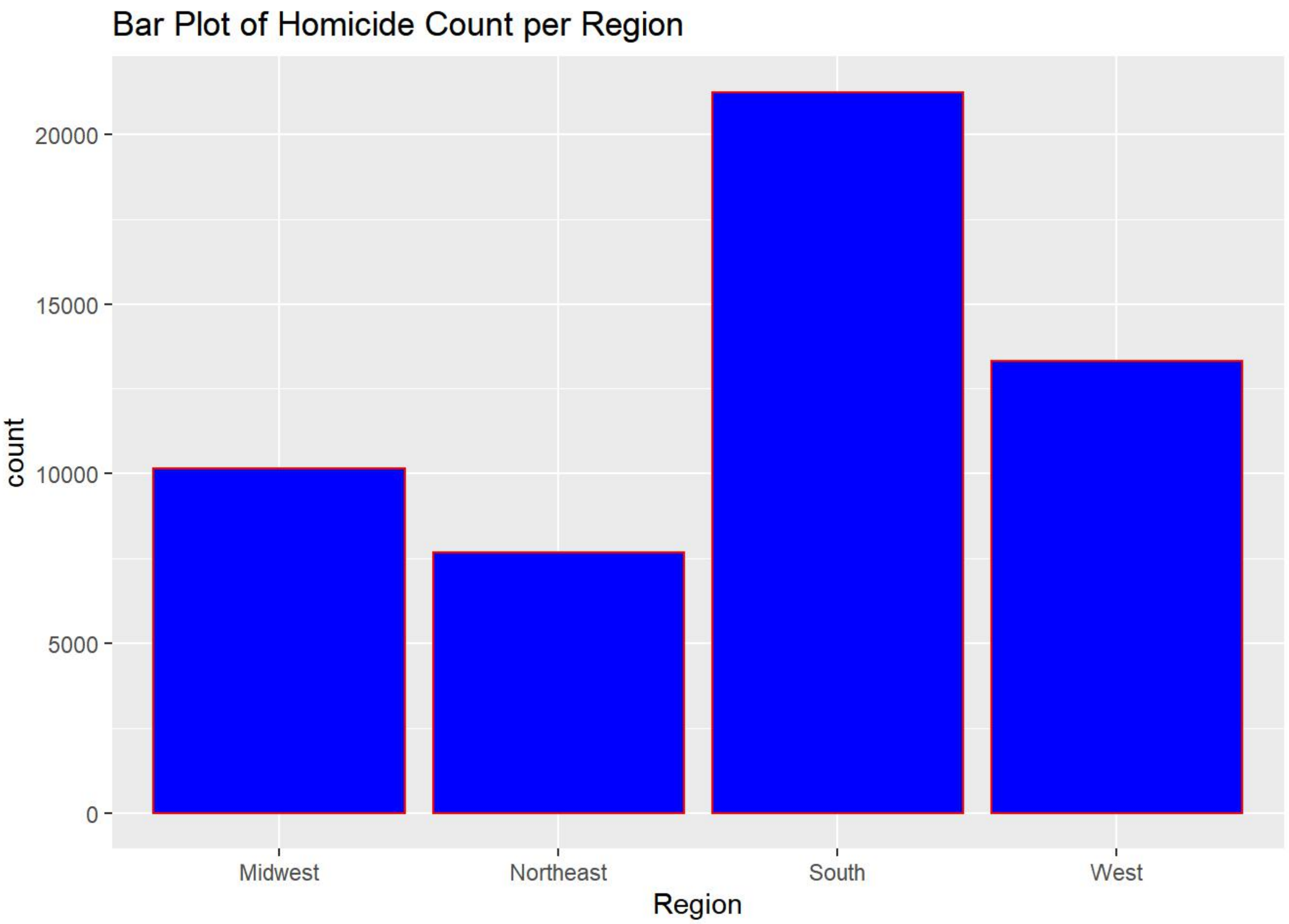
2.5.1 Region and homicides

To look at the relationship between region and homicides first we are going to subset the data so that we only have observations where there is at least one victim accounted for.

Hide

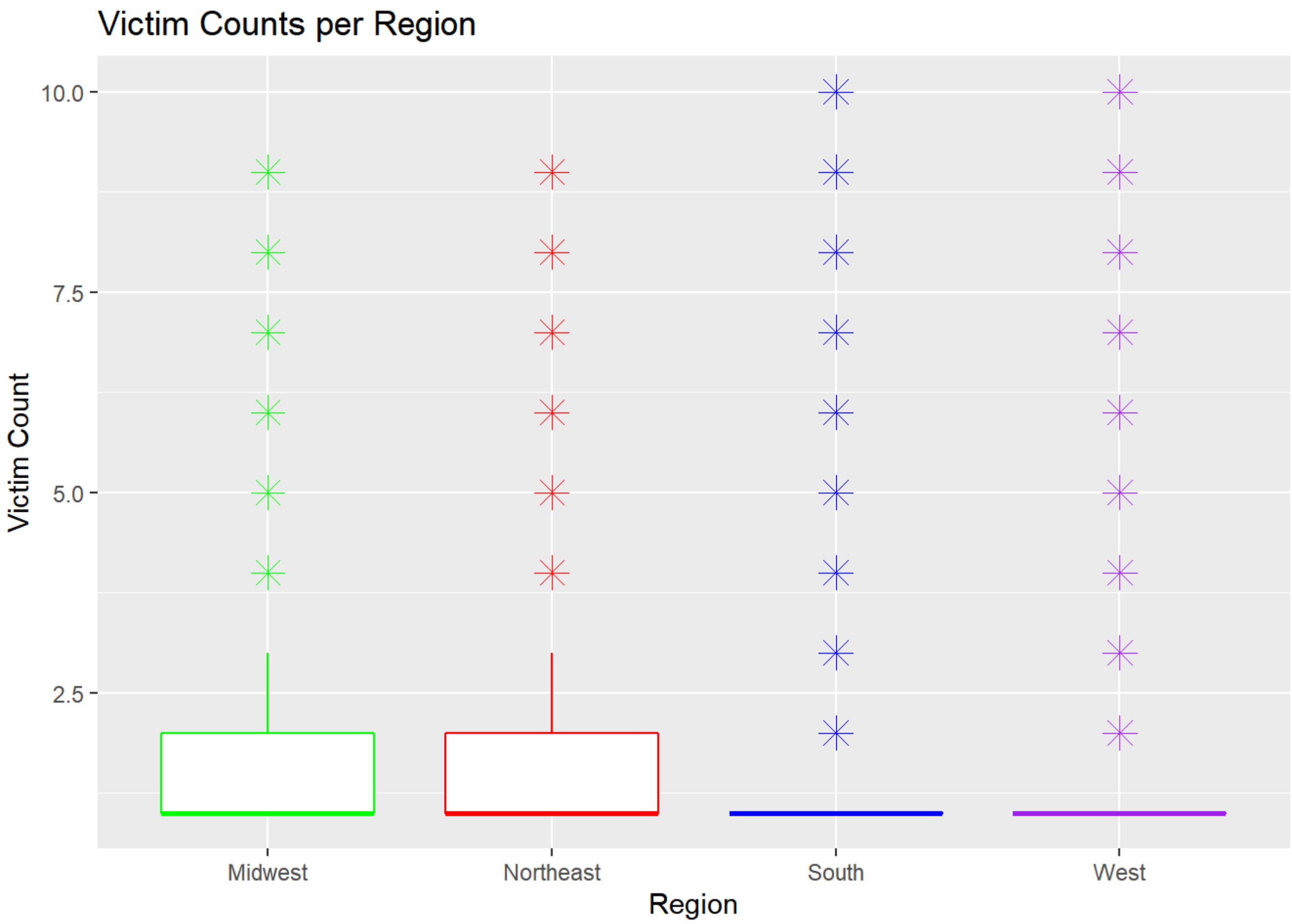
```
hom <- subset(crime, Victim.Count > 0)
hom$Region <- factor(hom$Region)

loadPkg("ggplot2")
ggplot(hom, aes(x=Region)) +
  geom_bar(col="red",fill="blue") + labs(title = "Bar Plot of Homicide Count per Region")
```



Hide

```
ggplot(hom, aes(x=Region, y=Victim.Count)) +
  geom_boxplot( colour=c("green","red","blue", "purple"), outlier.shape=8, outlier.size=4) +
  labs(title="Victim Counts per Region", x="Region", y = "Victim Count")
```

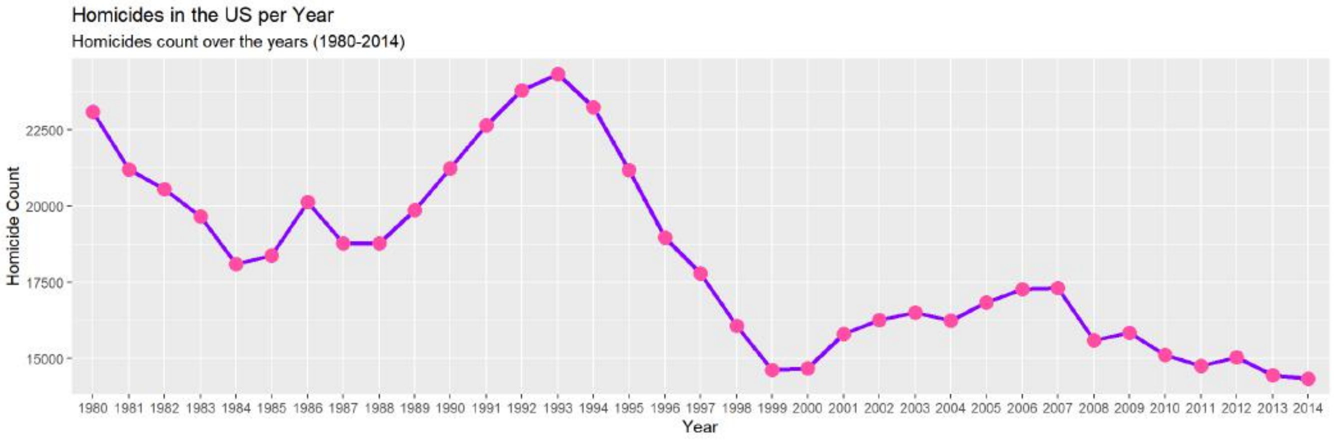


The graph above very clearly shows that the most homicides happen in the South region while the Northeast Region has the lowest count. Despite this we still need to account for population differences before drawing any final conclusions.

2.5.2 Year and solving cases

Hide

```
total <- table(crime$Year)
total <- as.data.frame(total)
names(total) <- c("Year", "Freq")
ggplot(total, aes(x=Year, y=Freq, group=1)) +
  geom_line(size=1.3, color="#8800ff") + geom_point(color="#ff4ea4", size=4) +
  labs(title="Homicides in the US per Year",
       subtitle="Homicides count over the years (1980-2014)",
       x="Year",
       y="Homicide Count")
```



The graph above shows that the homicide rate declines from 1980 to 2014 although it hits a high in 1993. The data spans over a large period of time so we are going to break the years up into groups of 5 to see trends more easily. We are also removing any NA values.

Hide

```
hom$Year <- cut(hom$Year, breaks = seq(1980, 2015, by = 5))
nas <- subset(hom, is.na(Year))
hom2<-anti_join(hom,nas)
```

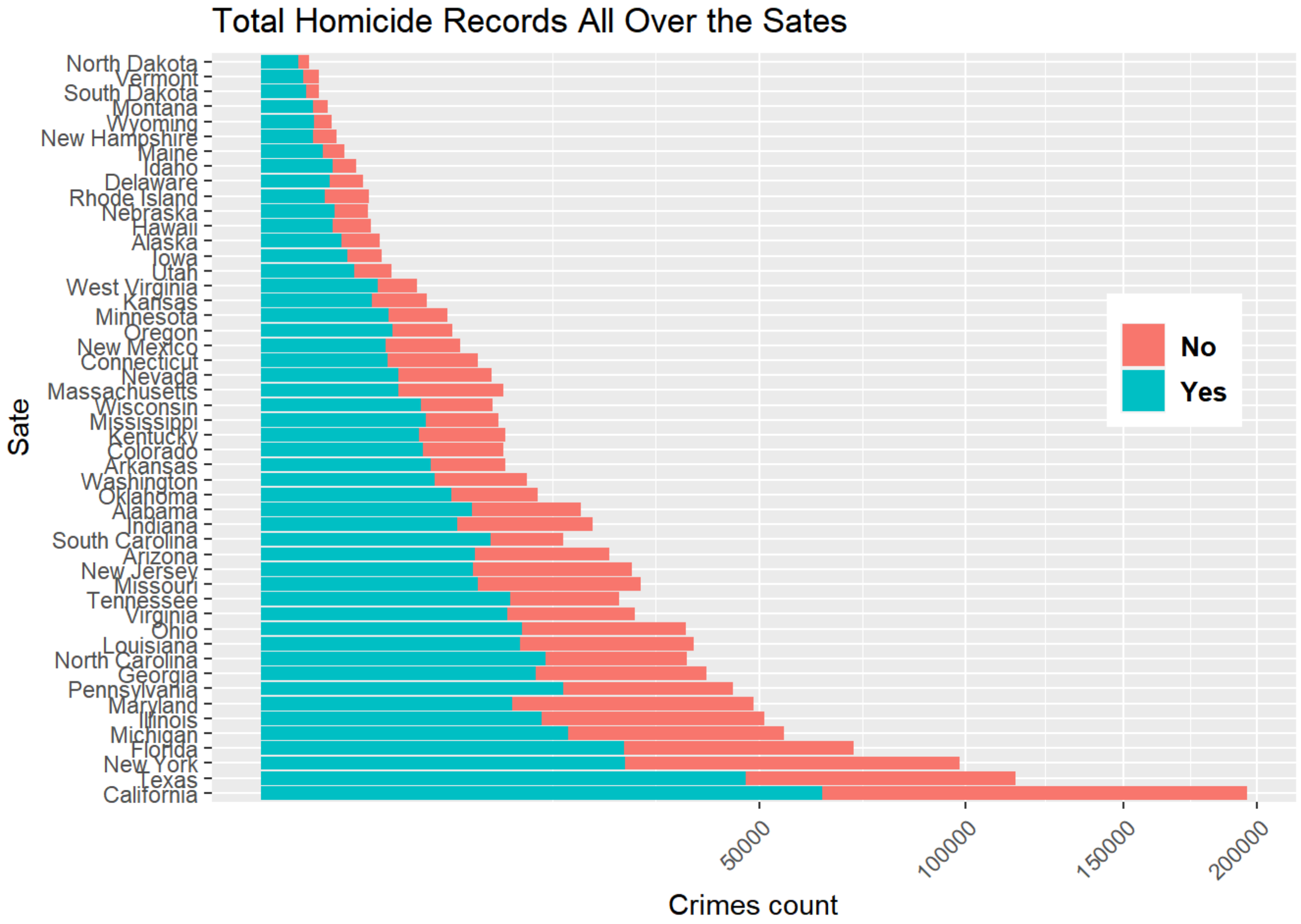
2.5.3 States and solving cases

Now we would like to find out which states are having more homicides and how many of them were solved and unsolved. After that, we'd like to focus on the top five states with most cases as they were dealing with more frequent homicides. Their efficiency would reflect more by the state's overall crime condition.

Hide

```
library(ggplot2)
state_data <- as.data.frame(table(crime$Crime.Solved, crime$State))
names(state_data) <- c("Crime.Solved", "State", "Freq")

ggplot(state_data, aes(reorder(State, -Freq), Freq, fill = Crime.Solved)) +
  geom_bar(stat="identity") +
  scale_y_sqrt() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.justification=c(1,0),
        legend.position=c(0.95,0.5),
        legend.text = element_text(size = 10, face = "bold"),
        legend.title = element_blank())+
  coord_flip()+
  labs(title="Total Homicide Records All Over the Sates",
       x="Sate",
       y="Crimes count",
       fill="Cases Solved")
```



If we look into the bottom of the graph we can see that the last few states with the most homicide cases are having almost double amount of cases compared to all other states in total count.

Top 5 states with most homicides and solved cases :

1. California
2. Texas
3. New York
4. Florida
5. Michigan

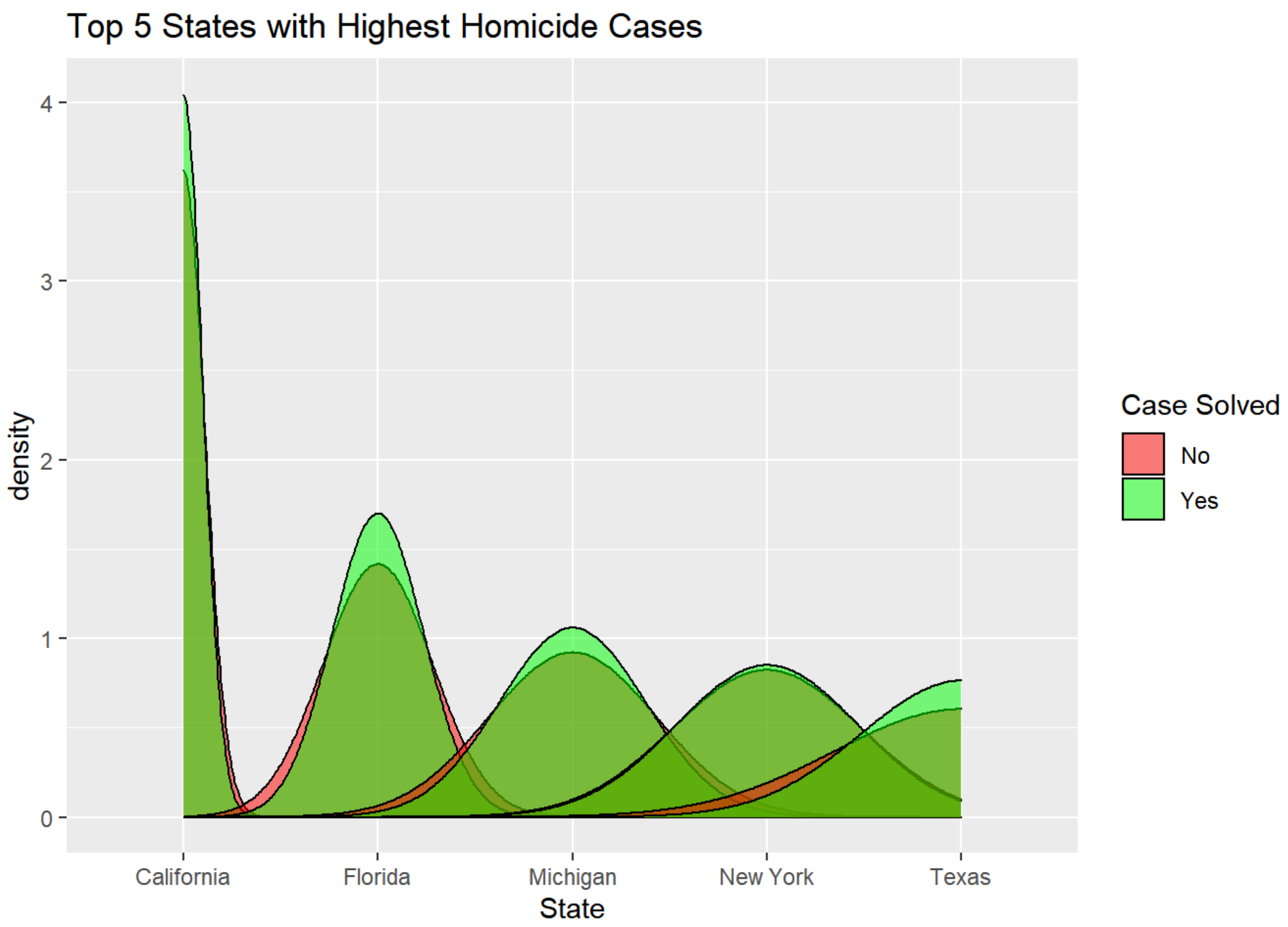
Most of these states have a the most eventful cities of the whole United States. There is Silicon Valley and San Francisco in California with fastest-growing industries of healthcare, construction, technology, hospitality, and agriculture. Texas has Dallas with technology, financial services and defense. There is New York City, one of the most famous city and city of dreams of every people around the world. It's Often referred to as the 'Big Apple', this vibrant city is known for the city that never sleeps. It's full of the cultures, business, and one of the media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports, and is the most photographed city in the world. It would be very interesting to find how the states with such big cities are controlling homicides.

Now, lets find out how efficient they are about solving cases.

Hide

```
crime$Crime.Solved  = factor(crime$Crime.Solved )
top51 <- subset(crime, State=="California")
top52 <- subset(crime, State=="Texas")
top53 <- subset(crime, State=="New York")
top54 <- subset(crime, State=="Florida")
top55 <- subset(crime, State=="Michigan")

top5 <- rbind(top51, top52)
top5 <- rbind(top5, top53)
top5 <- rbind(top5, top54)
top5 <- rbind(top5, top55)
ggplot(top5, aes(x=State, fill=Crime.Solved))+
  geom_density(alpha=.5)+
  scale_fill_manual(breaks = c("No", "Yes"),
                    values=c("red", "green"))+
  labs(title="Top 5 States with Highest Homicide Cases",
     fill="Case Solved")
```



From this density plot its clear that each of these 5 states have more solved cases than unsolved. Which is convincing but lets find out their success rates.

Success rate of the states:

California: 63.552%

Texas: 76.359%

New York: 54.104%

Florida: 71.351%

Michigan: 66.838%

Among the top five states with highest homicide cases, **Texas** can be said one of the best states at solving cases as it has the most success rate of 76%, higher than the other five. **Florida** is the second best among them with 71% rate of solving cases out of the total homicides. **California** and **Michigan** both have their success rate above 60%. But in **New York**, the solved and unsolved cases seems to be almost same. It is only around 50% of solved cases. Considering it has the most influential cities in the world, the success rates were expected much higher.

2.5.4 Weather type and homicides

Now we would like to find out if our data set can show some connection between the number of homicides and the weather temperature. But the American state is located in three climatic zones therefore it cannot be associated with a single distinctive landscape. There are the great plains affected by desert climate, California luxuriates in soft and warm. Whereas there are Mediterranean weather conditions and people of snow-covered Alaska are able to admire the aurora. Initially, we wanted to observe this relationship by season type. But the temperature is so versatile in all over the USA that it cannot be generalized that way. That’s why we are sub grouping the temperature data for better observations.

For our observation, we are sub-setting temperature by 4 categories. They are:-

Freezing: Temperature < 31F

Cold: Temperature between 31F and 68F

Warm: Temperature between 68F and 82F

Hot: Temperature > 82F

Hide

```
#### Subsetting temperature by value
df_freezing <- subset(crime, temp <= 31 )
df_cold <- subset(crime, temp > 31 & temp <= 68)
df_warm <- subset(crime, temp > 68 & temp <= 82 )
df_hot <- subset(crime, temp > 82)
# Categorizing the subsets to Freezing, Cold, Warm, Hot
df_freezing$Weather.Type <- "Freezing"
df_cold$Weather.Type <- "Cold"
df_warm$Weather.Type <- "Warm"
df_hot$Weather.Type <- "Hot"
# joining everything to one dataframe
total_df <- rbind(df_freezing, df_cold)
total_df <- rbind(total_df, df_warm)
total_df <- rbind(total_df, df_hot)
total_df <- total_df[order(total_df$Record.ID),]
#str(total_df)
```

Hide

```
analysis_df <- subset(total_df, select = c(City, State, Year, Month, Incident, Crime.Type, Crime.Solved, Relationship, Weapon, temp, Weather.Type))
#str(analysis_df)

analysis_df$Crime.Solved  = factor(analysis_df$Crime.Solved )
analysis_df$Weather.Type = factor(analysis_df$Weather.Type )
```

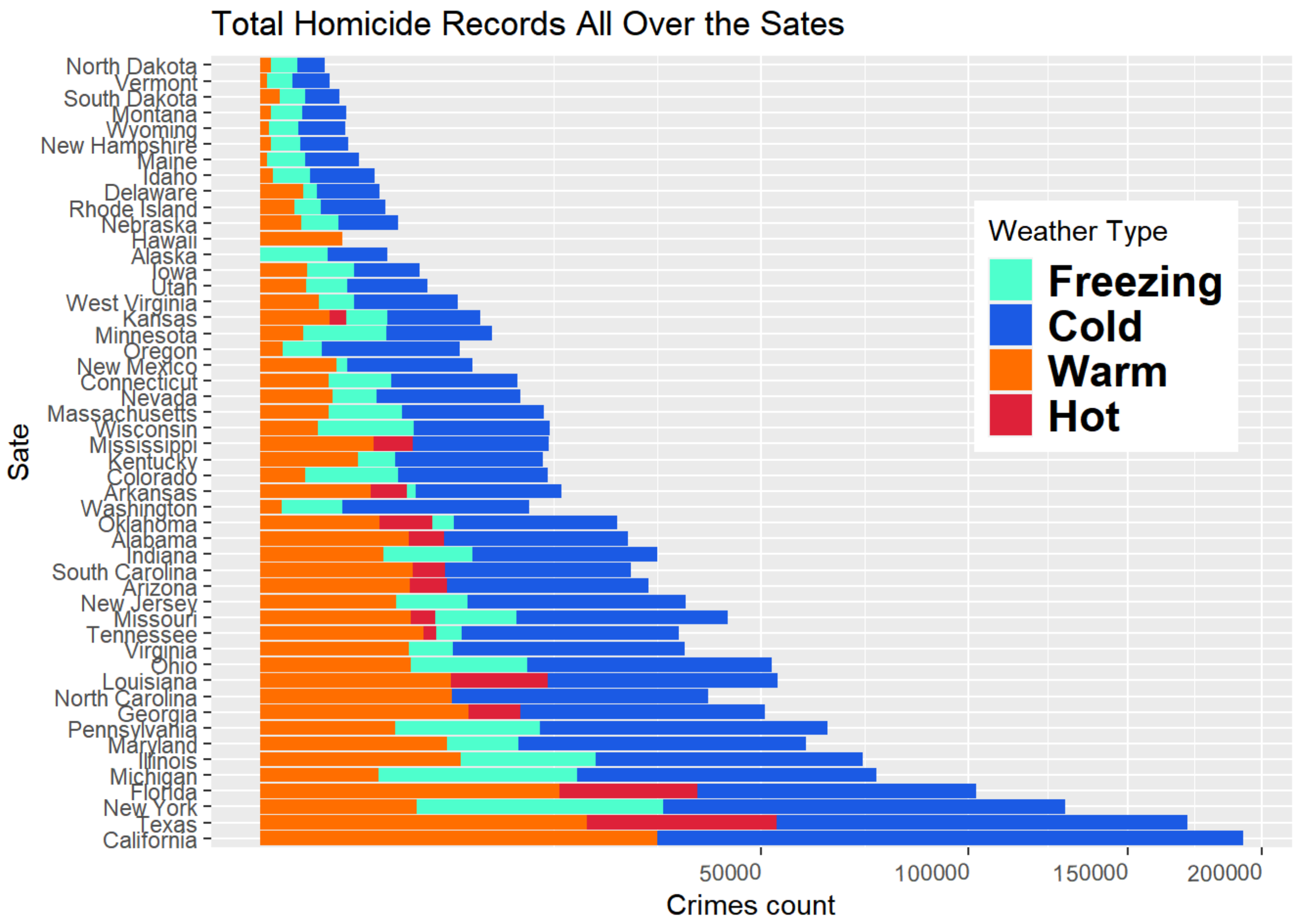
Hide

```
table1 <- table(analysis_df$State)
names(table1) <- c("State", "Total.Incidents")
```

Hide

```
library(ggplot2)
state_data <- as.data.frame(table(analysis_df$Weather.Type, analysis_df$State))
names(state_data) <- c("Weather.Type", "State", "Freq")

ggplot(state_data, aes(reorder(State, -Freq), Freq, fill = Weather.Type)) +
  geom_bar(stat="identity") +
  scale_y_sqrt() +
  scale_fill_manual(breaks = c("Freezing", "Cold", "Warm", "Hot"),
                       values=c("#4effcd", "#1b5ae4", "#ff6e00", "#DE2139")) +
  theme(axis.text.x = element_text(angle = rel(1.5), hjust = 1),
        legend.justification=c(1,0),
        legend.position=c(0.95,0.5),
        legend.text = element_text(size = rel(1.5), face = "bold"))+
  coord_flip()+
  labs(title="Total Homicide Records All Over the Sates",
       x="Sate",
       y="Crimes count",
       fill="Weather Type")
```



Hide

```
#ggsave("myplot.png")
```

Here, California is having the highest homicide incidents and majority of them are during cold weather. The second highest numbers of incidents happened in Texas and here as well the highest crimes happened during the cold weather and that continued to happen even in the hot temperature with prominent significance.

Top 5 states in most homicide cases are:

1. California
2. Texas
3. New York
4. Florida
5. Michigan

Top 5 states in least homicide cases are:

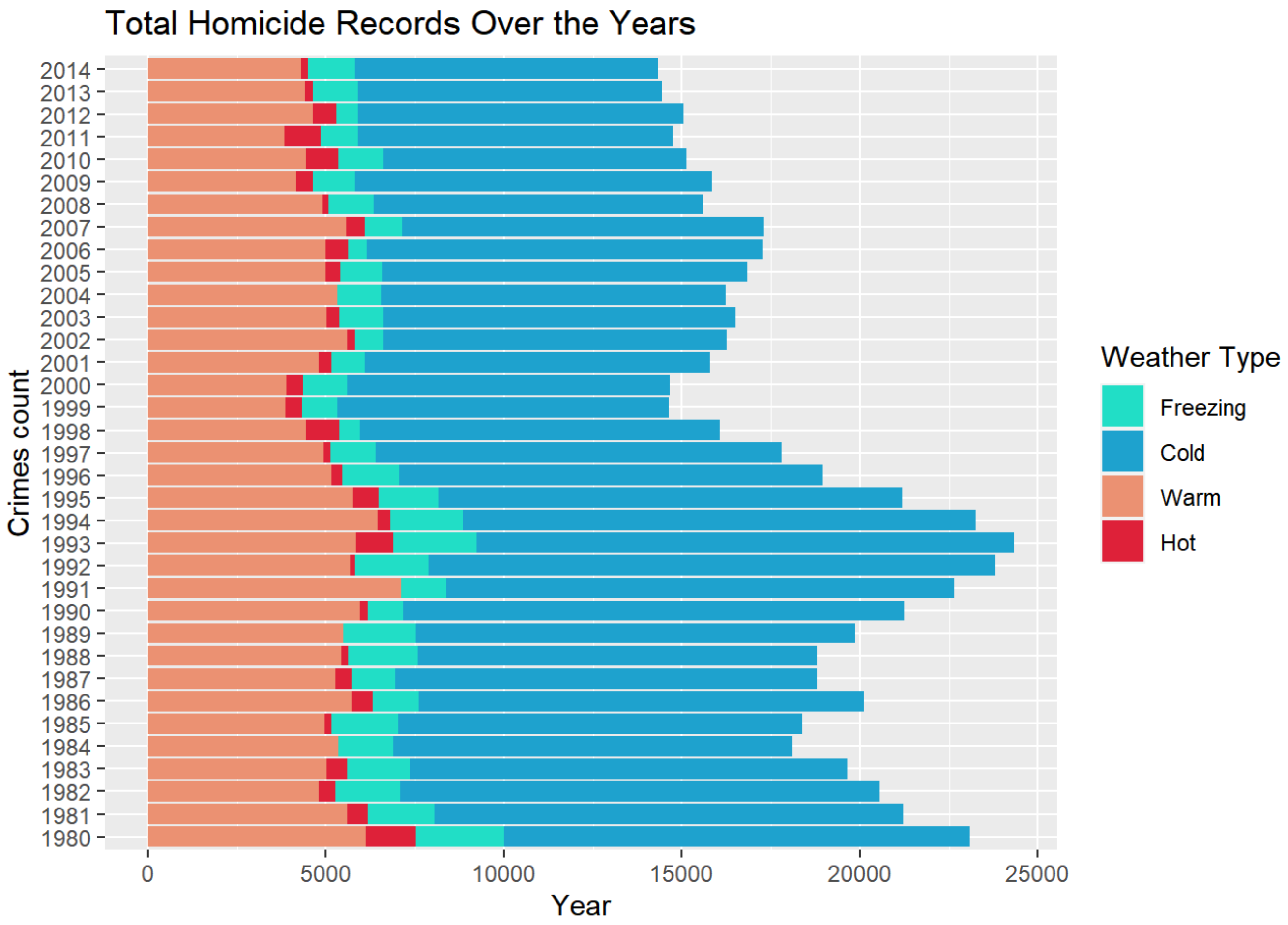
1. North Dakota
2. Vermont
3. South Dakota
4. Montana

5. Wyoming

Now let’s observe the homicides by years and weather types.

Hide

```
ggplot(analysis_df, aes(x=as.factor(Year), fill=Weather.Type )) +
  geom_bar( ) +
  scale_fill_hue(c = 30) +
  coord_flip()+
  scale_fill_manual(breaks = c("Freezing", "Cold", "Warm", "Hot"),
                    values=c("#21dec6", "#1EA2CE", "#EB9172", "#DE2139"))+
  labs(title="Total Homicide Records Over the Years",
     x="Crimes count",
     y="Year",
     fill="Weather Type")
```



Here also we can see most cases were during the cold weather over the years between 1980 and 2014. Along with that, even in the extreme temperatures the homicides continued to happen. Overall, it seems in warm temperature there are least cases of homicides.

It can be also observed that, there’s a significant drop of cases after 1993. After that it started to decrease gradually till the end of the century. The homicide cases in the 21st century were comparatively lower than the 20th. Maybe it could be the technological advancement that 21st century, people were more focused on the changes than committing homicides.

2.6 Tests and Modeling

2.6.1 Weapon and Type of relationship between victim and perpetrator

The first hypothesis test we will generate is for exploring the relationship between weapon type and victim/perpetrator relationship. Since these two variables are both factor variables and we are testing whether they are either independent or related, it seems appropriate to use a **Chi-Squared test of independence**.

Hide

```
#chi-squared test of independence
contable = table(homicides$Relationship_Type, homicides$Weapon); contable
```

After creating a contingency table between the two variables, we are now ready to run a Chi-Squared test of independence to examine the relationship, or lack thereof, between weapon type and victim/perpetrator relationship.

Hide

```
chitest = chisq.test(contable); chitest
chitest$statistic
chitest$p.value
chitest$parameter
xkabledply(chitest$observed)
xkabledply(chitest$expected)
xkabledply(chitest$residuals)
#reject null hypothesis; there is statistically significant association between relationship type and weapon used
```

The resultant statistic from the Chi-Squared test is 3885.253 with a p-value of 0. With such a small p-value that lies well below our threshold of $p < 0.05$, we reject the null hypothesis and conclude that there is a statistically significant association between relationship type and weapon used.

2.6.2 Region and homicides

Due to the fact that Region is a categorical variable and Victim Count is a numerical variable, the most appropriate test to choose here is ANOVA.

Hide

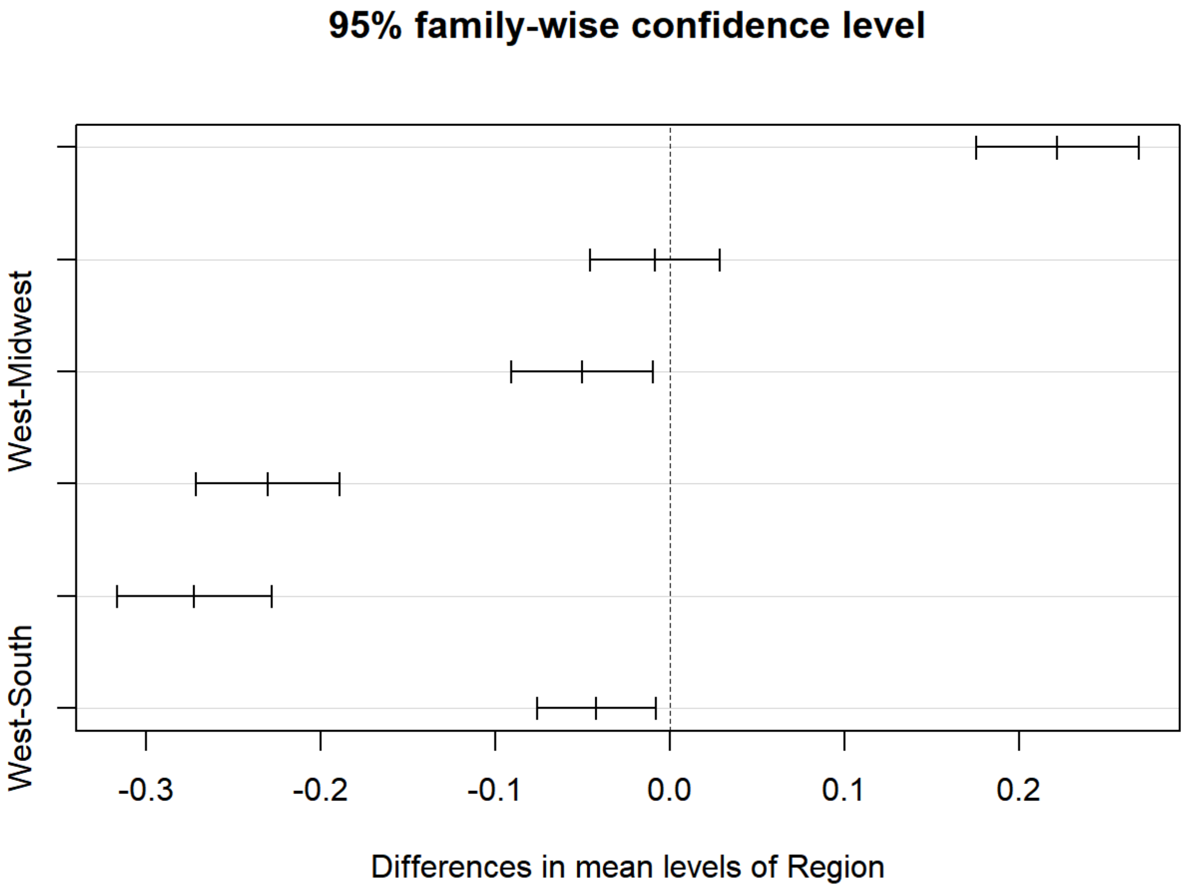
```
anovatest <- aov(Victim.Count ~ Region, data = hom)
xkabledply(anovatest)
anovasummary<- summary(anovatest)
```

The ANOVA test shows us that our p-value is very small at 1.081^{-59} so we can reject the null hypothesis that region and victim count are not dependent. Since we are rejecting the null hypothesis we can look at Tukey HSD.

Hide

```
tukey <- TukeyHSD(anovatest)
tukey

plot(tukey)
```



Hide

```
viccount<- table(hom$Region, hom$Victim.Count)
regtablechi <- chisq.test(viccount)
```

Tukey HSD reveals that the South-Midwest pair is the only one with no significance at the .05 level.

We can also take a look at a table showing the expected values versus the observed values, which also helps reveal that we have the highest victim counts in the South and Northeast regions.

Observed Victim Count for Each Homicide per Region

	1	2	3	4	5	6	7	8	9	10
Midwest	7408	1566	656	280	132	63	16	18	20	0
Northeast	5286	1263	532	190	90	63	56	36	160	0
South	15944	3263	1083	384	186	90	64	36	50	150

	1	2	3	4	5	6	7	8	9	10
West	10112	2064	576	230	102	70	32	54	60	10

Expected Victim Count for Each Homicide per Region

	1	2	3	4	5	6	7	8	9	10
Midwest	7513	1581	552	210	98.9	55.5	32.6	27.9	56.2	31.0
Northeast	5677	1195	417	159	74.7	41.9	24.6	21.1	42.5	23.4
South	15716	3308	1155	440	206.8	116.0	68.1	58.4	117.6	64.9
West	9844	2072	723	275	129.6	72.7	42.7	36.6	73.7	40.6

2.6.3 Year and solving cases

To begin testing the relationship between the year of the homicide and whether or not the case has been solved we can use a chi-squared test because we are dealing with two categorical variables.

Hide

```
solvedtable <- table(hom2$Crime.Solved, hom2$Year)
tablechi <- chisq.test(solvedtable)
tablechi
tablechi$p.value
```

The Pearson’s Chi-square test reveals that we have a very small p-value of 2.914^{-21}, which is less than .05 so we can reject the null hypothesis that year and crime solve status are not dependent. We can also take a look at the tables showing the Expected and Observed values for Victim Count per Region.

Observed Amount of Cases Solved vs Unsolved

	(1980,1985]	(1985,1990]	(1990,1995]	(1995,2000]	(2000,2005]	(2005,2010]	(2010,2015]
No	1965	1655	2721	1867	2152	2058	1515
Yes	4964	5283	6248	4724	5483	5678	4514

Expected Amount of Cases Solved vs Unsolved

	(1980,1985]	(1985,1990]	(1990,1995]	(1995,2000]	(2000,2005]	(2005,2010]	(2010,2015]
No	1899	1902	2459	1807	2093	2121	1653
Yes	5030	5036	6510	4784	5542	5615	4376

2.6.4 Weather type and homicides

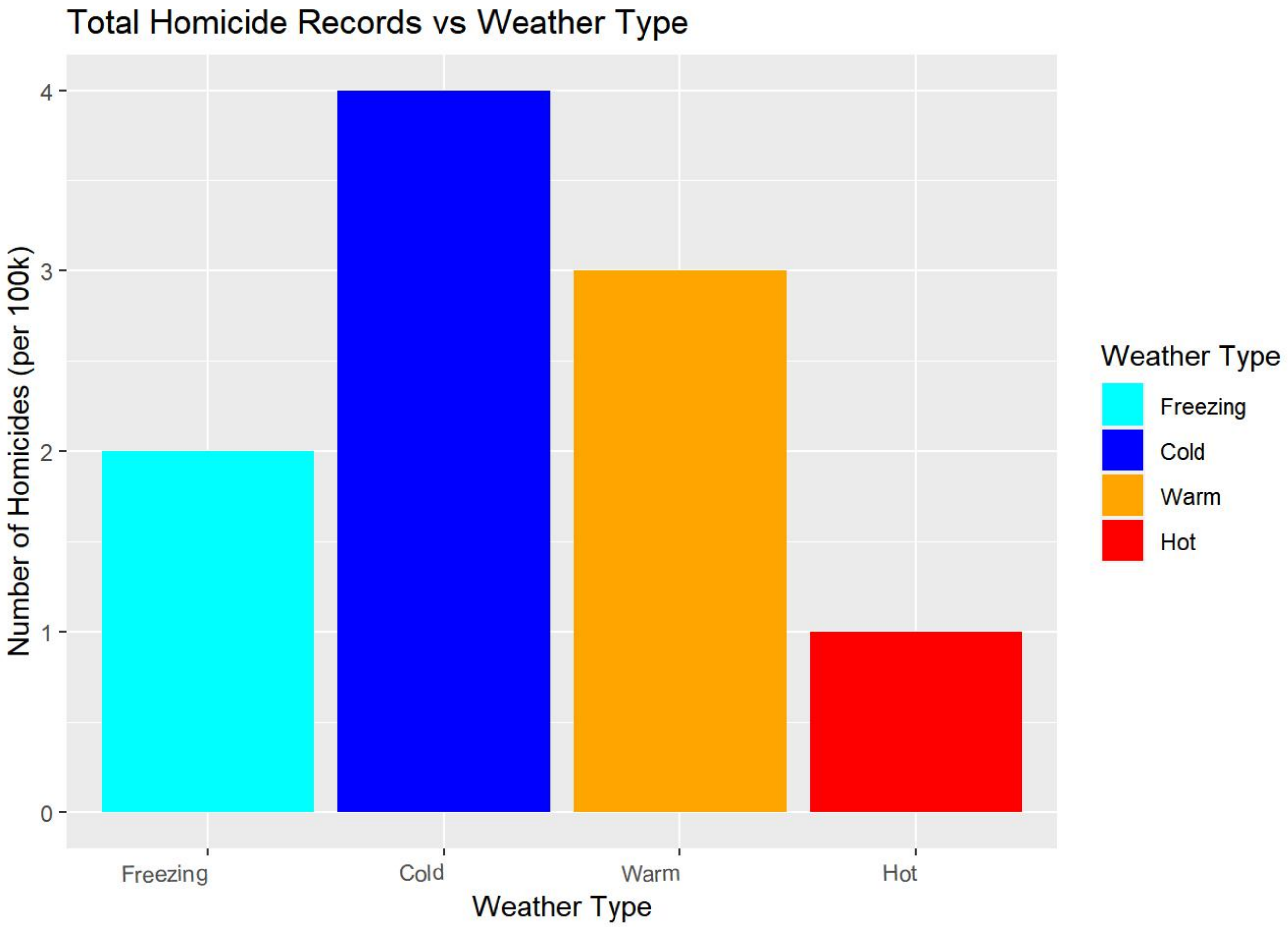
Now let’s find out if the years of incidents have any correlation with case solving or not.

Hide

```
weatherdata <- table(analysis_df$Weather.Type)
weatherdf <- data.frame(weatherdata)
weatherdf$Freq <- factor(weatherdf$Freq)
```

Hide

```
ggplot(weatherdf, aes(y=as.numeric(Freq), x=factor(Var1, level = c("Freezing","Cold","Warm","Hot")), fill=factor(Var1))) +
  geom_bar(stat='identity') +
  scale_fill_manual(breaks = c("Freezing", "Cold", "Warm", "Hot"),
                    values=c("cyan", "blue", "orange", "red"))+
  theme(axis.text.x = element_text(angle = rel(1.5), hjust = 1))+
  labs(title="Total Homicide Records vs Weather Type",
      x="Weather Type",
      y="Number of Homicides (per 100k)",
      fill="Weather Type")
```



Overall, mostly the incidents had happened during cold and warm weather.

To find out the correlation between count of homicides and weather type, we are doing a chi-squared test on weather type and count of homicides with a null hypothesis that Weather Type and number of homicides are independent.

Hide

```
monthlycrimedata <- table( analysis_df$State,analysis_df$Year, analysis_df$Month, analysis_df$Weather.Type)
monthlycrimedf <- as.data.frame(monthlycrimedata)
monthlycrimedf$Freq <- factor(monthlycrimedf$Freq)
#str(monthlycrimedf)
```

Hide

```
contable=table(monthlycrimedf$Var4, monthlycrimedf$Freq)

chitest = chisq.test(contable)
chitest
```

```
##
##  Pearson's Chi-squared test
##
## data:  contable
## X-squared = 22954, df = 1029, p-value <2e-16
```

Given this p-value of 0 is less than the alpha of 0.05, we reject the null hypothesis that Weather Type and number of homicides are independent. We conclude that there is evidence that the two variables are dependent (i.e., that there is an association between the two variables).

2.7 Findings

Hide

```
loadPkg("corrplot")
corrplot(chitest$residuals, is.cor = FALSE)
```

Hide

```
#red = fewer homicides than expected for relationship/weapon combo, blue = more homicides than expected
```

An intuitive way to visualize the results of the Chi-Squared test is to use a heatmap to examine the residuals of the expected vs. actual counts of the observations for each relationship type and weapon combination. In the heat map above, the red coloring indicate fewer homicides than expected (negative residual) for the relationship type and weapon, while the blue indicates a greater number of homicides than expected. The notable results here lie in the columns for 'Family Member' and 'Stranger/Unknown'. For the 'Family Member' relationship type, we note that we see significantly more observations than expected for 'Suffocation', 'Drowning', and 'Blunt Object'. On the other hand, we see significantly more 'Firearm' homicides than expected for the 'Stranger/Unknown' relationship type.

Returning to the question of if there is a region with more homicides, we can now adjust for population to see if the bar graph results can still stand. We will use population data for each region (United States Census Bureau). Homicides in the South make up 40.557% of the homicides in our data set compared to the approximately 37.2% of the population the region makes up. Homicides in the Northeast make up 14.65% compared to the approximately 17.8% of the population the region makes up. Homicides in the Midwest make up 19.389% the approximately 21.5% of the population the region makes up. Homicides in the West make up 25.403% the approximately 23.4% of the population the region makes up.

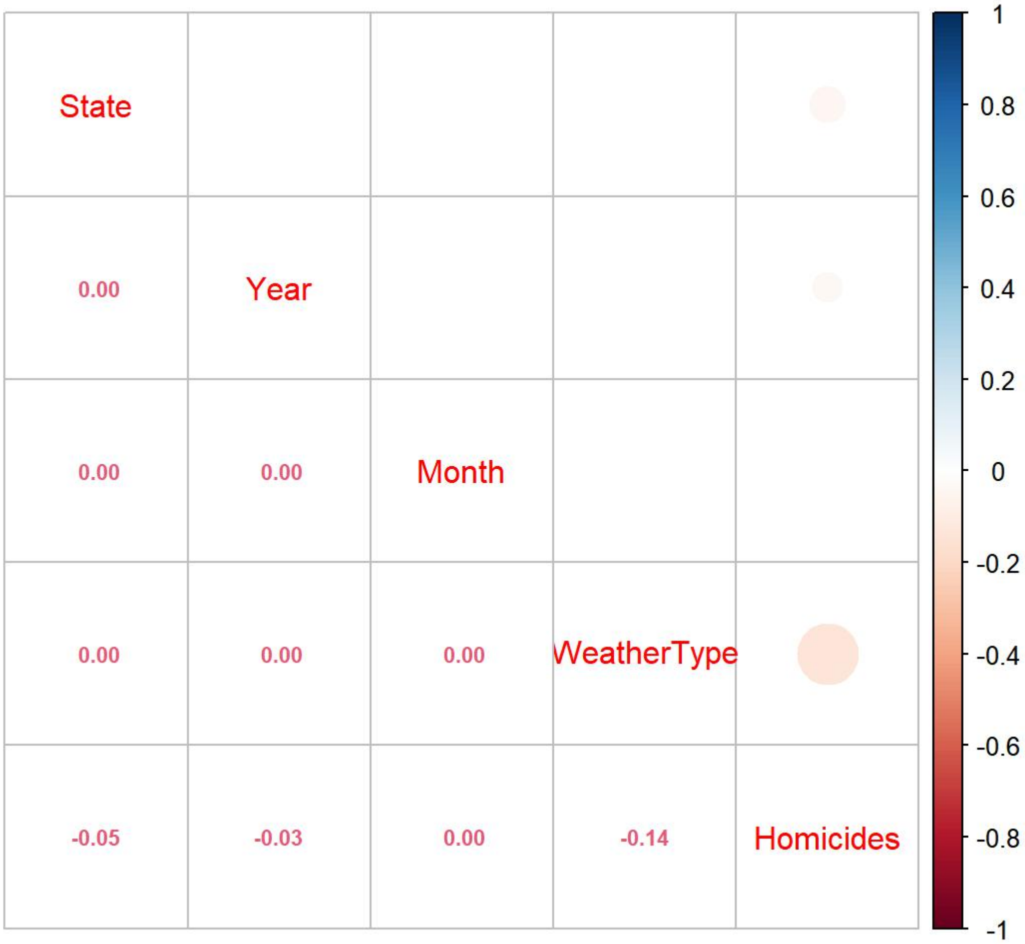
From 1980 to 1985 71.641 % of homicides were solved. From 1985 to 1990 76.146 % of homicides were solved. From 1990 to 1995 69.662 % of homicides were solved. From 1995 to 2000 71.673 % of homicides were solved. From 2000 to 2005 71.814 % of homicides were solved. From 2005 to 2010 73.397 % of homicides were solved. From 2010 to 2015 74.871 % of homicides were solved.

Now, we are going to check if there is any relationship between weather type and number of homicides by generating a correlation matrix.

Correlation between incident, year and weather type:

Hide

```
numericdata <- data.frame(monthlycrimedf)
numericdata$Var1 <- cbind(as.numeric(numericdata$Var1))
numericdata$Var2 <- as.numeric(numericdata$Var2)
numericdata$Var3 <- as.numeric(numericdata$Var3)
numericdata$Var4 <- as.numeric(numericdata$Var4)
numericdata$Freq <- as.numeric(numericdata$Freq)
colnames(numericdata)[0:5] = c("State","Year","Month","WeatherType","Homicides")
cormatrix = cor(numericdata)
corrplot.mixed(cormatrix,lower.col = "#e25978", number.cex = .7)
```



It can be observed that, th number of Homicides and Weather type has a very small negative relations. Which means as the temperature increases, number of homicide incidents decreases by 14%.

Overall, there isn’t that much of relationship between number of homicides and the weather type that we has assumed from the beginning. As USA has a very diverse climate, our data set couldn’t give that much insights of it. Also, as we had to use monthly average data of the state on analyze, it cannot precisely give the incident day temperature.

3 Conclusion

These are the final remarks on our observations all together,

1. Weapon used and perpetrator/victim relationship: For our first question pertaining to the possible relationship between the weapon used in a homicide and the relationship between the victim and perpetrator, we were able to conclude through the use of a Chi-Squared test that the two variables are associated. We saw more observations than expected for homicides involving family members committed through drowning, suffocation, and blunt objects. These weapons are very intimate, personal ways of taking another person's life. On the contrary, we saw more observations than expected for homicides between strangers (or unknown relationship) committed via firearm. Using a firearm is a non-intimate, ranged method of committing a homicide.

2.With a great portion of the data containing unknown information, the relationships between victim and perpetrator are comparatively vague in general. However, we can conclude that:

- The victim and perpetrator are likely from the same racial and/or age group
- Certain weapons contribute significantly to the variance in victim age, a relationship between the vitality of the weapons and the victim age is seen

3. After adjusting for population it becomes clear that the South region has the highest amount of homicides while the Northeast region has the lowest amount. Further investigation also reveals that the South and Northeast regions have homicides with the highest victim counts.
4. Conducting a chi-squared test showed that there is a relationship between they year of the homicides and whether or not the case has been solved. Looking at the first grouping of five years we can see that there are less cases solved than expected but as we progress through time there ends up being more cases solved than expected. There is an unexpectedly high number of cases solved from 1985 to 1990. One possible explanation for this could be that during that time there were a lot of serial killers who were getting arrested at different times for different crimes (Weiss).
5.
6.
7. Best state at solving cases: Texas is better at solving cases among the top 5 states with most homicide cases with having around 71% of solved cases among the total cases. Whereas, New York seems to be the worst among them. It could solve around only 50% of cases, even tough having one of the most influential big city in it.
8. Weather type and homicide: After our observations, it's significant that most homicide cases happened during the cold weathers and least during the extreme hot temperatures. It should be also keep in mind that overall USA has a cooler temperature, so the cases are supposed to be more at that type of weather. Over all, after our tests we can conclude that, there could be an association between weather type and number of homicides. As the weather temperature increases, number of homicides decreases by 14%.

4 References

Census Regions and Divisions of the United States. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf (https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf).

U.S. Census Bureau United States Population Growth by Region. (https://www.census.gov/popclock/print.php?component=growth&image=//www.census.gov/popclock/share/images/growth_1561939200.png (https://www.census.gov/popclock/print.php?component=growth&image=//www.census.gov/popclock/share/images/growth_1561939200.png))

Weiss, Debra Cassens. “Serial Killings Are Waning, Leading to Speculation about the Cause.” ABA Journal, https://www.abajournal.com/news/article/serial-killings-are-waning-leading-to-speculation-about-the-cause (https://www.abajournal.com/news/article/serial-killings-are-waning-leading-to-speculation-about-the-cause).