Homicides vs Educational Attainment

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(tidyr)
library(rmarkdown)
library(readr)
library(viridis)
```

Load data

Part 1: Data

Plan Phase of the PACE Framework

How are the observations in the sample collected?

The Kaggle dataset is based on the FiveThirtyEight project, which acquired gun-related death information from the Centers for Disease Control and Prevention's Multiple Cause of Death database. The original dataset only contained data from 2012 to 2014. This updated dataset includes data for additional years, from 2006 to 2011 and from 2015 to 2020.

The observations are likely independent of one another, as they were randomly sampled. However, it is impossible to confirm this with 100% certainty, as there may be some unknown factors that could have influenced the independence of the observations.

As the data collected is based on observations rather than a controlled experiment, it is not possible to infer causation.

Part 2: Research question

Plan Phase of the PACE Framework

Project Scope

The goal of this project is to investigate the correlation (if any) between the level of education of a homicide victim and their likelihood of being a victim of homicide. Specifically, we want to determine whether people with lower levels of education are more likely to be victims of homicide than people with higher levels of education.

Part 3: Exploratory data analysis

Analyze Phase of the PACE Framework

Investigate the Dataset

Initial investigations of the dataset suggest a correlation between lower educational attainment and a higher rate of homicides. As educational attainment increases, the homicide rate appears to decrease.

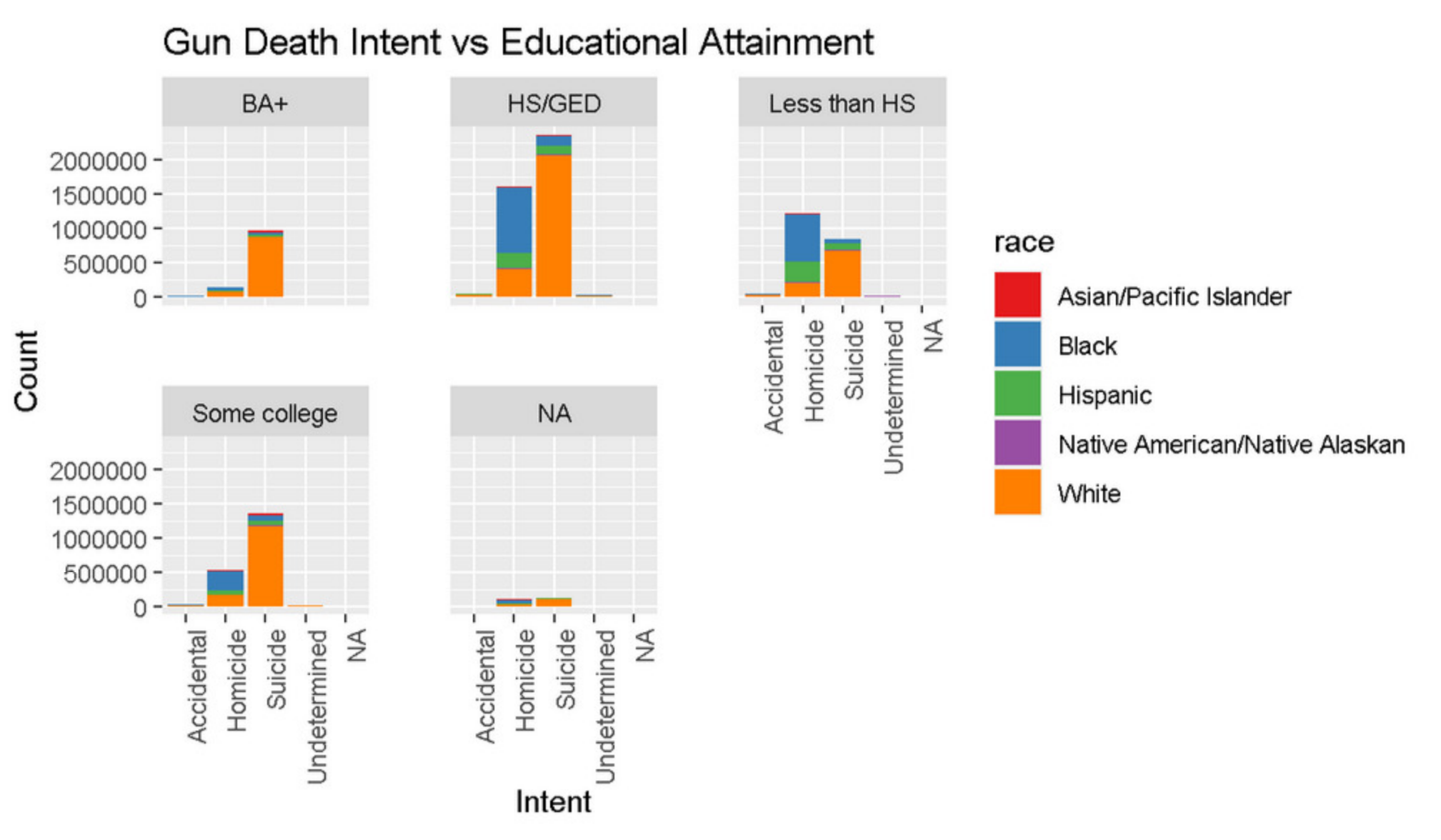
There are several factors that could contribute to this being accurate:

- Educational attainment is associated with socioeconomic status. People with lower educational attainment are more likely to live in poverty, which is a risk factor for violence. Poverty can lead to stress, frustration, and a lack of opportunities, which can increase the likelihood of violence.
- Educational attainment is associated with cognitive ability. People with lower educational attainment may have lower cognitive abilities, which can make it more difficult for them to solve problems peacefully. They may also be more impulsive and less able to control their emotions, which can increase the risk of violence.
- Educational attainment is associated with social norms. People with lower educational attainment may be more likely to live in communities where violence is more common. They may also be more likely to be exposed to violent media, which can normalize violence and make it more likely that they will engage in violence themselves.

```
# Initial investigation of key variables
table(gun_crime_data$race,
      gun_crime_data$education,
      gun_crime_data$intent)
```

```
## , ,  = Accidental
##
##
##                                    BA+  HS/GED Less than HS Some college
##   Asian/Pacific Islander           180     540          450          324
##   Black                            918   10458        13500         3528
##   Hispanic                         288    4068         7326         1728
##   Native American/Native Alaskan   108     738         1206          432
##   White                          10134   40122        27108        18990
##
## , ,  = Homicide
##
##
##                                    BA+  HS/GED Less than HS Some college
##   Asian/Pacific Islander         10008   21924        11952        11826
##   Black                          43866  956502       687870       285858
##   Hispanic                       10890  214344       307548        60012
##   Native American/Native Alaskan   972   15408        13518         5778
##   White                          85536  408780       205002       169236
##
## , ,  = Suicide
##
##
##                                    BA+  HS/GED Less than HS Some college
##   Asian/Pacific Islander         21906   20088         8298        20412
##   Black                          33732  144378        62784        80550
##   Hispanic                       27666  115542        96426        69282
##   Native American/Native Alaskan  2934   21726        12528        11466
##   White                         880074 2067786       672516      1173312
##
## , ,  = Undetermined
##
##
##                                    BA+  HS/GED Less than HS Some college
##   Asian/Pacific Islander           144     324          180          306
##   Black                            468    5958         5580         2034
##   Hispanic                         324    2718         4014         1206
##   Native American/Native Alaskan     0     666          648          234
##   White                           5886   22500        10854        11538
```

```
# Plot visualizing potentially relevant variables
ggplot(gun_crime_data, aes(x = intent, fill = race)) +
  geom_bar(stat = "count") +
  facet_wrap(~education) +
  labs(title = "Gun Death Intent vs Educational Attainment",
       x = "Intent",
       y = "Count") +
  scale_fill_brewer(palette = "Set1") +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1),
    panel.spacing = unit(1, "cm")
  )
```



Data Manipulation and Cleaning

```
# Keep columns for analysis
gun_crime_data_trimmed <- gun_crime_data %>%
  select(c("intent","education"))

# Remove all rows with irrelevant data for this analysis
rows_to_remove <- grepl("suicide|accidental|undetermined", gun_crime_data_trimmed$intent, ignore.case = TRUE)
gun_crime_data_trimmed <- gun_crime_data_trimmed[!rows_to_remove, ]

# Remove all rows with irrelevant data for this analysis
if (FALSE) {
gun_crime_data_trimmed <- gun_crime_data_trimmed[!grepl("Suicide", gun_crime_data_trimmed$intent), ]
gun_crime_data_trimmed <- gun_crime_data_trimmed[!grepl("Accidental", gun_crime_data_trimmed$intent), ]
gun_crime_data_trimmed <- gun_crime_data_trimmed[!grepl("Undetermined", gun_crime_data_trimmed$intent), ]
}

# Check for Missing Values
sum(is.na(gun_crime_data_trimmed$education))
```

```
## [1] 99594
```

```
sum(is.na(gun_crime_data_trimmed$intent))
```

```
## [1] 486
```

```
# Omit rows with missing values
if (FALSE) {
gun_crime_data_trimmed <- na.omit(df)
}

# Impute the missing rows with the mode.
mode <- names(which.max(table(gun_crime_data_trimmed$education)))
gun_crime_data_trimmed$education[is.na(gun_crime_data_trimmed$education)] <- mode
mode <- names(which.max(table(gun_crime_data_trimmed$intent)))
gun_crime_data_trimmed$intent[is.na(gun_crime_data_trimmed$intent)] <- mode

# Replace missing values with the most frequent value
if (FALSE) {
most_frequent_value <- levels(gun_crime_data_trimmed$intent)[which.max(table(gun_crime_data_trimmed$intent))]
gun_crime_data_trimmed <- gun_crime_data_trimmed %>%
mutate(intent = replace_na(intent, most_frequent_value))

most_frequent_value <- levels(gun_crime_data_trimmed$race)[which.max(table(gun_crime_data_trimmed$race))]
gun_crime_data_trimmed <- gun_crime_data_trimmed %>%
mutate(race = replace_na(race, most_frequent_value))

most_frequent_value <- levels(gun_crime_data_trimmed$education)[which.max(table(gun_crime_data_trimmed$education))]
gun_crime_data_trimmed <- gun_crime_data_trimmed %>%
mutate(education = replace_na(education, most_frequent_value))
}

table(gun_crime_data_trimmed$education,
      gun_crime_data_trimmed$intent)
```

```
##
##                Homicide
##   BA+            151362
##   HS/GED        1716696
##   Less than HS  1225980
##   Some college   532872
```

```
ggplot(gun_crime_data_trimmed, aes(x = intent)) +
  geom_bar(stat = "count") +
  facet_wrap(~education) +
  labs(title = "Homicide Victimization vs Educational Attainment",
       x = "",
       y = "Count") +
  scale_fill_brewer(palette = "Set1") +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    panel.spacing = unit(1, "cm"),
  )
```


```
ggsave("Homicides_vs_Educational Attaintment.png", width=6, height=8)
```

Conditions

Test Type

To analyze the data and determine if there is a correlation between the rate of homicide victimization and the level of educational attainment, a CHI-Square Independence Test will be used.

Conditions for the chi-square test:

1. Independence: Sampled observations must be independent.

- Random sample/assignment **Condition met;** *original location of data mentioned it was acquired through random sampling.*
- If sampling without replacement, n < 10% of population **Condition met.**
- Each case only contributes to one cell in the table *It is possible that individuals may have been a part of more than one group, such as attaining more than one educational group, such as HS/GED and College. This condition can be difficult to achieve during an* observational study.*

2. Sample size: Each particular scenario (i.e. cell) must have at least 5 expected cases **Condition met.**

Part 4: Modeling/Inference

Construct Phase of the PACE Framework

H0 (nothing going on): Homicide victimization and educational attainment are independent. Homicide rates do not vary by educational status.

HA (something going on): Homicide victimization and educational attainment are dependent. Homicide rates do vary by educational status.

```
# Create a contingency table of the data
contingency_table <- table(gun_crime_data_trimmed$intent, gun_crime_data_trimmed$education)

# Run a Chi-Square Independence Test on the table.
chi_square_result <- chisq.test(contingency_table)

# Extract the test statistics
chi_square <- chi_square_result$statistic
df <- chi_square_result$parameter

# Set the desired confidence level
confidence_level <- 0.95

# Calculate the critical value
critical_value <- qchisq(1 - (1 - confidence_level) / 2, df)

# Calculate the margin of error
margin_of_error <- sqrt(chi_square / sum(chi_square_result$observed)) * critical_value

# Calculate the lower and upper bounds of the confidence interval
lower_bound <- chi_square / (1 - margin_of_error)
upper_bound <- chi_square / (1 + margin_of_error)
```

Part 5: Prediction/Conclusion

Based on the p-value being much lower than the significance value of 0.05, it can be concluded that there is a correlation between the rate of homicide victimization and educational attainment. However, it is important to note that since this is an observational study and not an experiment, causation cannot be inferred.

```
# Print the confidence interval
cat("Confidence Interval:", lower_bound, "-", upper_bound, "\n")
```

```
## Confidence Interval: -308651.7 - 223465.9
```

```
# Print the Chi Square Test Result
print(chi_square_result)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  contingency_table
## X-squared = 1619358, df = 3, p-value < 2.2e-16
```

What Story does the Data Tell?

Execute Phase of the PACE Framework

Based on the data analysis, it appears that there is a connection between the level of education attained and the probability of becoming a victim of homicide. While there are several other factors that contribute to this conclusion, for the purposes of this analysis, this information should suffice.