# Exploring the Impact of Registered and Unregistered Guns on Public Safety

Derek Papierski

April 7, 2023

## Introduction

Gun ownership levels and their potential relationship with crime and safety have been a topic of interest and debate for many years. The central question revolves around whether higher levels of gun ownership contribute to increased crime rates or act as a deterrent, promoting safety. This research paper aims to examine the global landscape of gun ownership, crime, and safety indices, and investigate how these factors vary.

In this study, we begin by visually exploring the data through a series of graphs and numerical summaries that highlight the relationship between gun ownership and crime index across different regions. These visualizations offer a preliminary understanding of global patterns and potential correlations between gun ownership and crime. To further examine these patterns, we perform a bootstrap confidence interval analysis and a hypothesis test to assess if there are significant differences in crime levels between countries with high and low numbers of guns.

Finally, we fit and compare two regression models to evaluate the impact of regional factors and unregistered guns on crime indices. These models provide insights into the factors that may contribute to crime and the potential role of gun ownership in shaping crime and safety landscapes. By conducting this comprehensive analysis, we aim to provide a deeper understanding of the complex relationship between gun ownership, crime, and safety around the world.

***Note: This project seeks to maintain neutrality, avoiding any bias towards or against guns, and purely aims to understand the patterns present in the data.***

## External Requirements

```
# read in the data
df = read.csv("guns_dataset.csv")
```

```
# load libraries
library(ggplot2)
library(dplyr)
library(tidyr)
library(mosaic)
```

## Data Cleaning

In the dataset, the numbers in the unregistered guns and registered guns columns were reading in as characters instead of doubles because the numbers had commas in them in the excel sheet. To fix this, we simply

removed the commas in the columns and converted them back to numbers. There were also some NAs that were hard-coded as "N/A" and others were just empty strings. To fix this, we replaced the "N/A" values in the dataset with NA so that all the missing values in the dataset were all coded the same.

```
# remove commas from unreg_guns and reg_guns and convert them to numeric

df <- df %>%
  # remove commas
  mutate(Registered.firearms = gsub(",", "", Registered.firearms),
         Unregistered.firearms = gsub(",", "", Unregistered.firearms))
# convert to numeric
df <- df %>%
  mutate(Registered.firearms = as.numeric(Registered.firearms),
         Unregistered.firearms = as.numeric(Unregistered.firearms))
```

```
# replace "N/A" with NA

df[df=='N/A'] = NA
```

## Count and Remove NAs

When counting and removing the missing values in the dataset, we found that most of the missing values were from the Crime Index and Safety Index columns and the Registered and Unregistered guns columns. After deciding on the variables we want to work with, renaming them, and removing the rows that were incomplete, we ended up with a total of 104 rows.

```
# count and remove NAs

# see dimensions and summary
dim(df)
```

```
## [1] 230  12
```

```
summary(df)
```

```
##       Rank        Country.or.subnational.area
##  Min.   :  1.00   Length:230
##  1st Qu.: 58.25   Class :character
##  Median :115.50   Mode  :character
##  Mean   :115.50
##  3rd Qu.:172.75
##  Max.   :230.00
##
##  Estimate.of.civilian.firearms.per.100.persons    Region
##  Min.   :  0.000                               Length:230
##  1st Qu.:  2.100                               Class :character
##  Median :  5.900                               Mode  :character
##  Mean   :  9.808
##  3rd Qu.: 13.525
##  Max.   :120.500
##
```

```
##   Subregion         Population.2017
##  Length:230         Min.   :1.000e+03
##  Class :character   1st Qu.:6.060e+05
##  Mode  :character   Median :5.522e+06
##                     Mean   :3.269e+07
##                     3rd Qu.:2.049e+07
##                     Max.   :1.388e+09
##
##  Estimate.of.firearms.in.civilian.possession Computation.method
##  Min.   :      100                           Min.   :1.000
##  1st Qu.:    34500                           1st Qu.:2.000
##  Median :   245000                           Median :2.000
##  Mean   :  3777075                           Mean   :2.113
##  3rd Qu.:  1142500                           3rd Qu.:2.000
##  Max.   :393347000                           Max.   :3.000
##  NA's   :3
##  Registered.firearms Unregistered.firearms  Crime.Index     Safety.Index
##  Min.   :     48     Min.   :       50     Min.   :13.78   Min.   :16.42
##  1st Qu.:  16470     1st Qu.:    56168     1st Qu.:32.77   1st Qu.:44.33
##  Median : 142149     Median :   253302     Median :45.81   Median :54.19
##  Mean   : 734740     Mean   :  5029361     Mean   :44.89   Mean   :55.11
##  3rd Qu.: 541922     3rd Qu.:   816775     3rd Qu.:55.67   3rd Qu.:67.23
##  Max.   :9700000     Max.   :392273257     Max.   :83.58   Max.   :86.22
##  NA's   :94          NA's   :94            NA's   :95      NA's   :95
```

```
# see all NAs in each variable
colSums(is.na(df))
```

```
##                                          Rank
##                                             0
##                       Country.or.subnational.area
##                                             0
## Estimate.of.civilian.firearms.per.100.persons
##                                             0
##                                        Region
##                                             0
##                                     Subregion
##                                             0
##                                Population.2017
##                                             0
##   Estimate.of.firearms.in.civilian.possession
##                                             3
##                               Computation.method
##                                             0
##                             Registered.firearms
##                                            94
##                           Unregistered.firearms
##                                            94
##                                  Crime.Index
##                                            95
##                                Safety.Index
##                                            95
```

```r
# see number of complete rows
sum(complete.cases(df)) # 104
```

```
## [1] 104
```

```r
# see number of incomplete rows
sum(!complete.cases(df)) # 126
```

```
## [1] 126
```

```r
# take only columns we are interested in
selected_variables = select(df, Country.or.subnational.area,
                            Estimate.of.civilian.firearms.per.100.persons,
                            Region, Population.2017, Registered.firearms,
                            Unregistered.firearms, Crime.Index, Safety.Index)

# rename columns
selected_variables <- selected_variables %>%
  rename(country = Country.or.subnational.area,
         guns_per_100 = Estimate.of.civilian.firearms.per.100.persons,
         region = Region,
         population = Population.2017,
         reg_guns = Registered.firearms,
         unreg_guns = Unregistered.firearms,
         crime_index = Crime.Index,
         safety_index = Safety.Index)

# count missing rows
sum( !complete.cases(selected_variables) )
```

```
## [1] 126
```

```r
# remove missing rows
selected_variables = selected_variables[complete.cases(selected_variables), ]

# confirm there are no missing values
colSums( is.na(selected_variables) )
```

```
##      country guns_per_100       region   population     reg_guns   unreg_guns
##            0           0            0            0            0            0
##  crime_index safety_index
##            0           0
```

```r
knitr::kable(head(selected_variables),
             caption = 'Dataset Sample',
             col.names = c('Country', 'Guns per 100', 'Region', 'Population',
                           'Reg Guns', 'Unreg Guns', 'Crime Idx', 'Safety Idx'))
```

Table 1: Dataset Sample

| | Country | Guns per 100 | Region | Population | Reg Guns | Unreg Guns | Crime Idx | Safety Idx |
|---|---|---|---|---|---|---|---|---|
| 1 | United States | 120.5 | Americas | 326474000 | 1073743 | 392273257 | 48.16 | 51.84 |
| 5 | Montenegro | 39.1 | Europe | 626000 | 103536 | 141464 | 41.10 | 58.90 |
| 6 | Serbia | 39.1 | Europe | 6946000 | 1186086 | 1532914 | 38.29 | 61.71 |
| 7 | Canada | 34.7 | Americas | 36626000 | 2081442 | 10626558 | 42.95 | 57.05 |
| 8 | Uruguay | 34.7 | Americas | 3457000 | 605313 | 592687 | 51.44 | 48.56 |
| 9 | Cyprus | 34.0 | Asia | 839000 | 154327 | 130673 | 32.12 | 67.88 |

## Dataset Description

The dataset we are using for the project contains the estimated number of civilian guns per capita by country and the crime and safety index. This dataset combines information from two sources: the estimated number of civilian guns per capita by country is sourced from Small Arms Survey and came from the report "Civilian Firearms Holdings, 2017", while the crime and safety index numbers are obtained from Numbeo. Each row in the dataset represents a country or a subnational area, providing information about various variables for that specific location. After removing rows with missing values, the dataset contains 104 rows, having removed 126 rows from the original 230. The variables of interest include the categorical variable `Country.or.subnational.area`, which represents a country or subnational area, the categorical variable `Region`, which represents the geographic region where the country is located, and numerical variables such as `Estimate.of.civilian.firearms.per.100.persons` (ratio representing firearms per 100 persons), `Population.2017` (people or inhabitants), `Registered.firearms` (firearms), `Unregistered.firearms` (firearms), `Crime.Index` (a dimensionless value quantifying the level of crime), and `Safety.Index` (a dimensionless value quantifying the level of safety). This dataset is observational in nature, meaning that associations and correlations can be inferred, but not causation.

## Data Transformation

We created several secondary data frames that we used later for calculating statistics or making various plots. These secondary data frames include a regional gun data data frame where we calculated the regional numbers and grouped the data by region, a data frame that holds the 10 countries with the highest crime index, two data frames separating low gun countries and high gun countries (using the median guns per capita as the threshold), and a data frame where we removed one of the extreme outliers.

```
# recode country and region as factors
selected_variables <- selected_variables %>%
  mutate(country = as.factor(country),
         region = as.factor(region))
```

```
# regional gun data sub dataframe
# for multiple graphs

guns_data_by_region <- selected_variables %>%
  group_by(region) %>%
  summarise(region_avg_guns_per_100 = mean(guns_per_100),
            region_reg_guns = sum(reg_guns),
            region_unreg_guns = sum(unreg_guns),
            region_avg_crime_index = mean(crime_index),
            region_avg_safety_index = mean(safety_index))
```

```r
# top 10 countries with highest crime index score sub dataframe
# for graph 1 in part A

top_10_countries_crime <- selected_variables %>%
  arrange(desc(crime_index)) %>%
  head(10)
```

```r
# make two subset groups of high and low gun countries
# use the median guns_per_100 as cutoff for groups (median=9.8)
# for bootstrap confidence interval and hypothesis test

# decide on the threshold for groups
# see what the median guns_per_100 is
# median(selected_variables$guns_per_100) # 9.8

# low gun countries subset
low_gun_countries <- subset(selected_variables,
                            guns_per_100 <= median(guns_per_100),
                            select = c(country,
                                       guns_per_100,
                                       crime_index))
# high gun countries subset
high_gun_countries <- subset(selected_variables,
                            guns_per_100 > median(guns_per_100),
                            select = c(country,
                                       guns_per_100,
                                       crime_index))
#low_gun_countries
#high_gun_countries
```

```r
# remove the unreg_guns outlier (United States)
# for regression model

remove_us_outlier <- selected_variables %>%
  filter(!(country == "United States"))
```

## Exploratory Data Analysis: Descriptive Statistics and Visualizations

### Summary Statistics

To start off our EDA, let's take a broad look at the summary statistics of our numerical variables.

```r
# table of summary statistics of the numerical variables

# empty data frame to store summary stats
summary_stats <- data.frame()
# vector with all the numerical variables
numerical_vars <- c('guns_per_100',
                    'population',
                    'reg_guns',
                    'unreg_guns',
                    'crime_index',
```

```
                    'safety_index')
# loop through numerical variables vector
for (var in numerical_vars) {
  # calculate stats of the variable
  var_stats <- favstats(selected_variables[[var]])
  # add new column to summary stats df
  var_stats <- cbind(variable = var, var_stats)
  # append the summary stats to summary_stats df
  summary_stats <- rbind(summary_stats, var_stats)
}

# print the summary statistics table
knitr::kable(summary_stats,
             caption = "Summary Statistics",
             digits = 1,
             col.names = c('Variable', 'Min', 'Q1', 'Median', 'Q3',
                           'Max', 'Mean', 'SD', 'n', 'NAs'))
```

Table 2: Summary Statistics

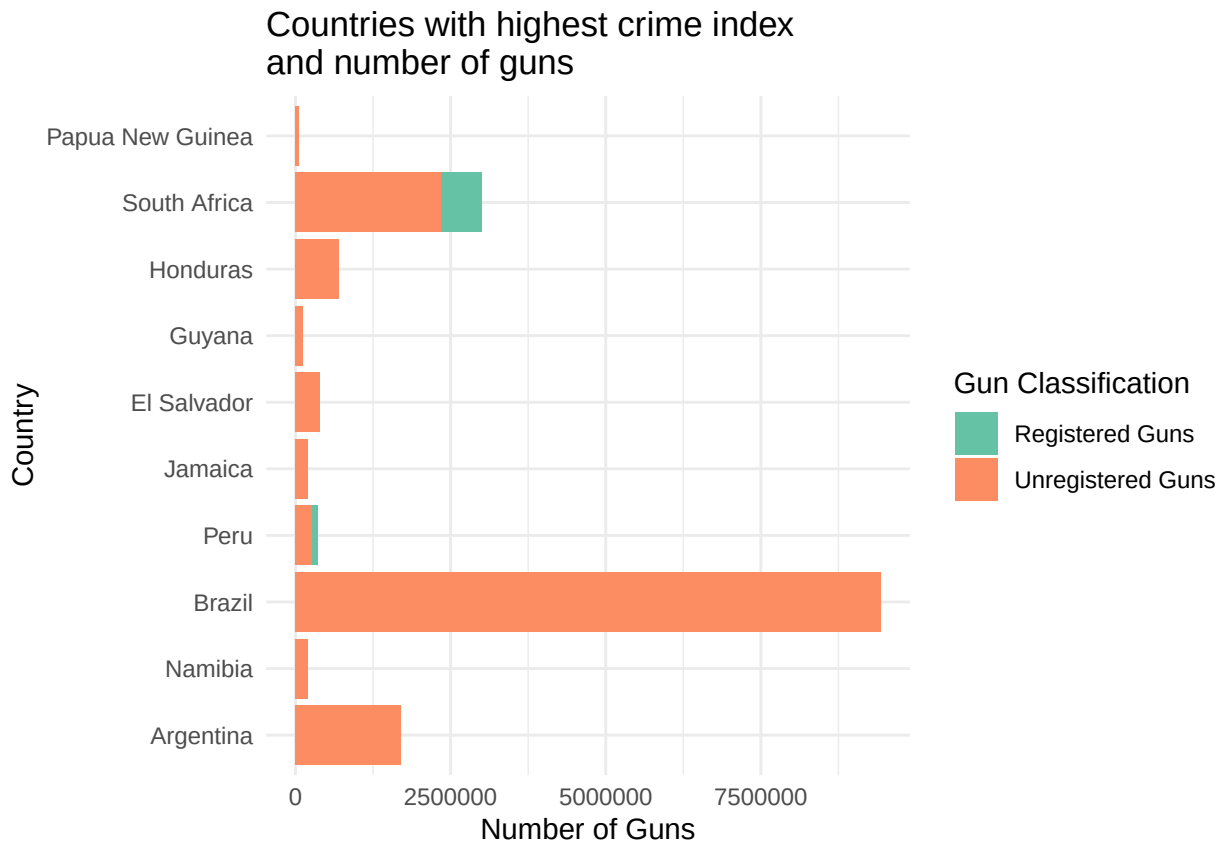
| Variable | Min | Q1 | Median | Q3 | Max | Mean | SD | n | NAs |
|---|---|---|---|---|---|---|---|---|---|
| guns_per_100 | 0.0 | 2.8 | 9.8 | 16.0 | 120.5 | 12.3 | 14.6 | 104 | 0 |
| population | 286000.0 | 4182500.0 | 10661000.0 | 41781250.0 | 1388233000.0 | 58705894.2 | 191947603.1 | 104 | 0 |
| reg_guns | 795.0 | 57278.2 | 202678.5 | 814421.2 | 9700000.0 | 933169.9 | 1765002.5 | 104 | 0 |
| unreg_guns | 3462.0 | 141940.8 | 302922.0 | 1335750.0 | 392273257.0 | 6525358.9 | 39167430.8 | 104 | 0 |
| crime_index | 15.9 | 33.7 | 44.9 | 54.4 | 81.2 | 44.8 | 14.0 | 104 | 0 |
| safety_index | 18.8 | 45.6 | 55.1 | 66.3 | 84.1 | 55.2 | 14.0 | 104 | 0 |

### Exploratory Visualizations

To assess the relationship between public safety and the presence of registered or unregistered guns, I wanted to visualize data from the ten countries with the highest crime index scores, specifically focusing on the numbers of registered and unregistered guns in each. The corresponding bar chart arranges these countries from highest to lowest crime index, juxtaposing this with the distribution of registered and unregistered guns.

```
# bar chart of 10 countries with highest crime index and number of guns plot

ggplot(top_10_countries_crime,
       aes(x = reorder(country, crime_index),
           y = crime_index)) +
  geom_bar(aes(fill = "Registered Guns",
               y = reg_guns),
           stat = "identity") +
  geom_bar(aes(fill = "Unregistered Guns",
               y = unreg_guns),
           stat = "identity") +
  coord_flip() +
  labs(x = "Country",
       y = "Number of Guns",
       title = "Countries with highest crime index \nand number of guns",
```

```
      fill = "Gun Classification") +
  scale_fill_manual(values = c("Registered Guns" = "#66c2a5",
                               "Unregistered Guns" = "#fc8d62")) +
  theme_minimal()
```

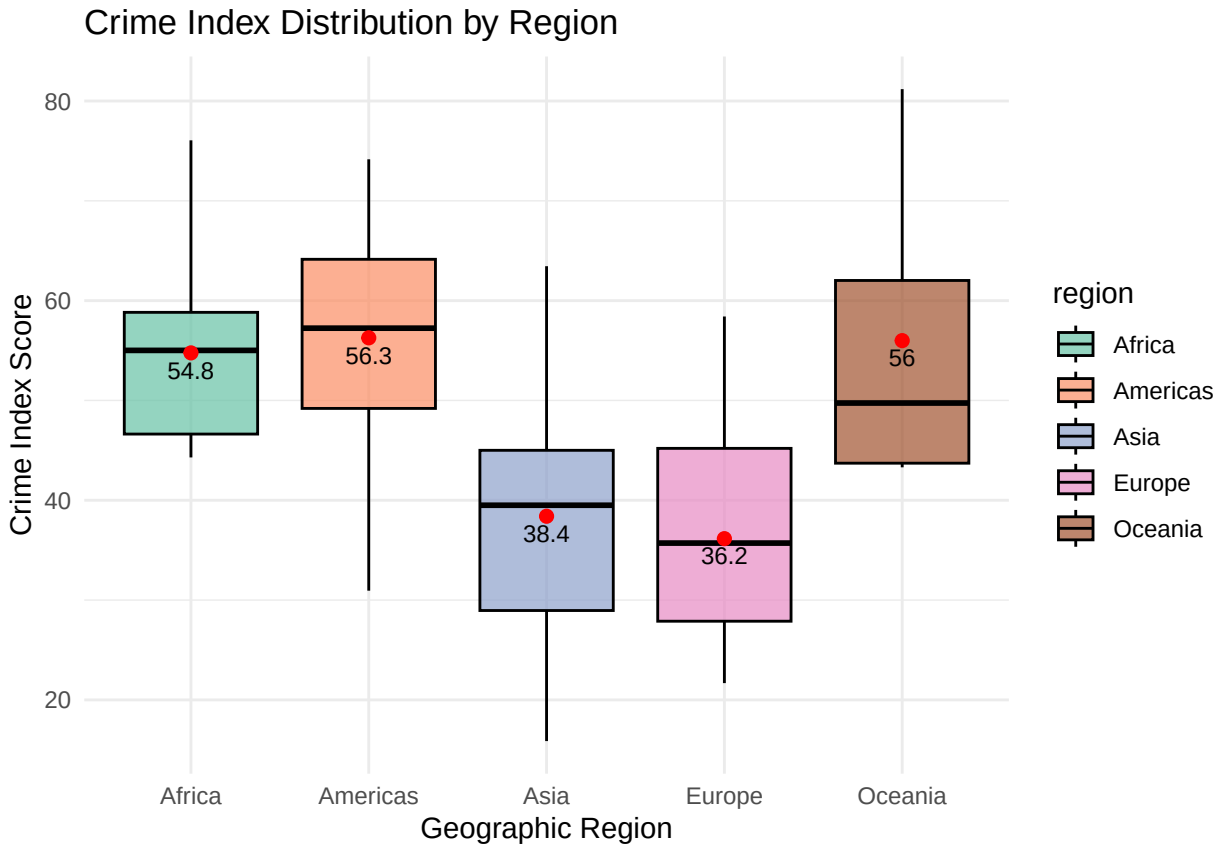



```
ggsave("plot1_highcrime.jpg")
```

In order to investigate if registered or unregistered guns is a bigger factor on public safety, I first wanted to see a bar chart of the 10 countries with the highest crime index score and see how many unregistered and registered guns were in that country. This bar chart shows the 10 countries with the highest crime indexes in order from highest to lowest and the number of registered and unregistered guns in that country. As we can see, a common theme is that a vast majority of guns in these countries are unregistered. Another interesting thing that we can see is that Papua New Guinea has the highest crime index score, but is the country with the least amount of guns compared to the others shown in the graph.

```
# box plot of crime index distribution by region

# find the means of each region to show on plot
means <- aggregate(crime_index ~ region, selected_variables, mean)

# make box plot
ggplot(data = selected_variables,
       aes(x = region,
           y = crime_index,
           fill = region)) +
```

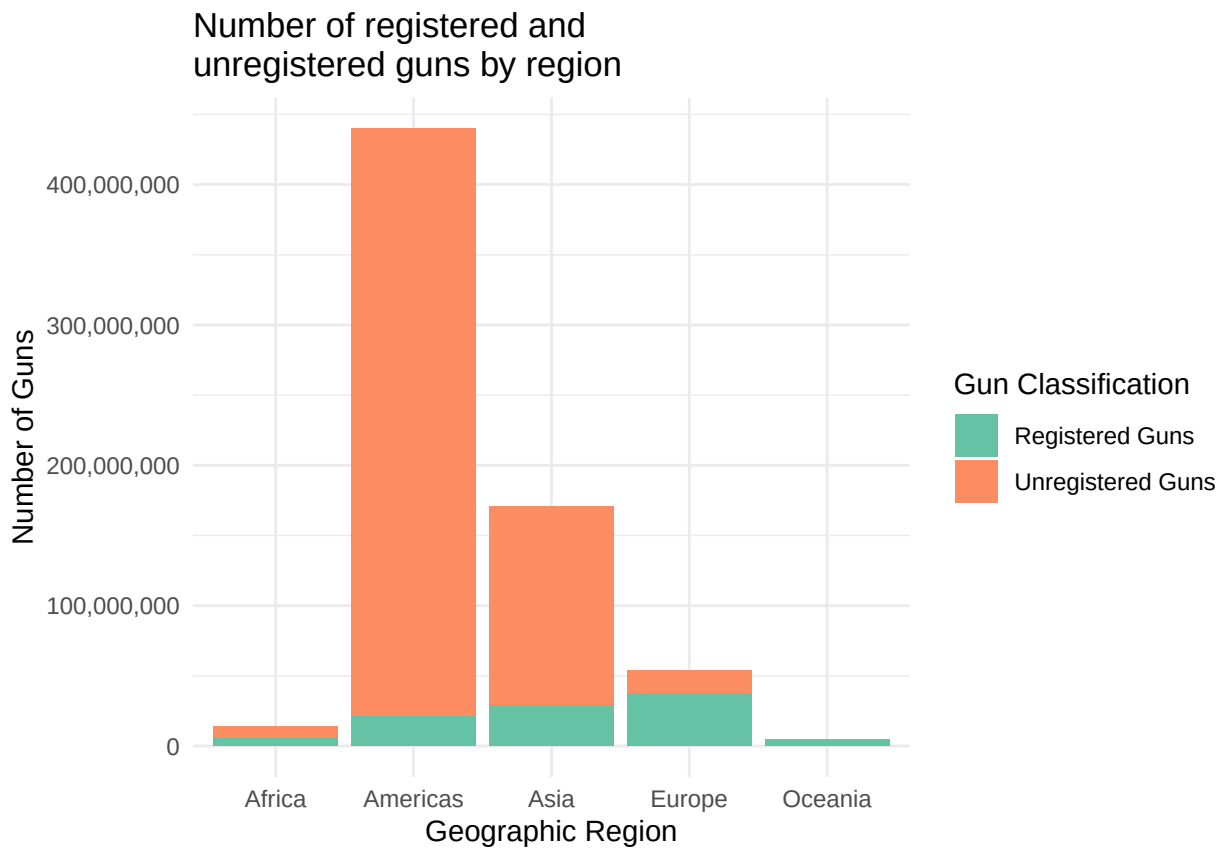
```
  geom_boxplot(color = "black",
               alpha = 0.7) +
  labs(title = "Crime Index Distribution by Region",
       x = "Geographic Region",
       y = "Crime Index Score") +
  scale_fill_manual(values = c("Africa" = "#66c2a5",
                               "Americas" = "#fc8d62",
                               "Asia" = "#8da0cb",
                               "Europe" = "#e78ac3",
                               "Oceania" = "sienna")) +
  # add mean crime index score to boxes
  stat_summary(fun = mean,
               color = "red",
               geom = "point",
               size = 2,
               show.legend = FALSE) +
  # add mean crime index score text for each box
  geom_text(data = means,
            aes(label = round(crime_index, 1),
                vjust = 1.5),
            size = 3) +
  theme_minimal()
```




```
ggsave("plot2_crimedist.jpg")
```

From the bar chart showing the top 10 countries with the highest crime index scores and the number of guns, we can see that many of the countries are in the Americas geographic region. To dig a little deeper into this, the box plot above illustrates the average crime index distribution by geographic region and their mean crime index scores. We can see that the Americas have the highest average crime index score with a mean of 56.3 and median of about 58, with Oceania being a close second. The plot also shows that Asia and Europe have very similar crime index scores and are two of the safest regions, having a mean of 38.4 and 36.2 respectively.

```
# stacked bar chart comparing number of guns across regions

ggplot(data = guns_data_by_region,
       aes(x = region)) +
  # bar chart for unregistered guns
  geom_bar(aes(y = region_unreg_guns,
               fill = "Unregistered Guns"),
           stat = "identity",
           position = "dodge") +
  # bar chart for registered guns
  geom_bar(aes(y = region_reg_guns,
               fill = "Registered Guns"),
           stat = "identity",
           position = "dodge") +
  # show numbers instead of scientific notation
  scale_y_continuous(labels = scales::number_format(big.mark = ",")) +
  labs(x = "Geographic Region",
       y = "Number of Guns",
       title = "Number of registered and \nunregistered guns by region",
       fill = "Gun Classification") +
  scale_fill_manual(values = c("Registered Guns" = "#66c2a5",
                               "Unregistered Guns" = "#fc8d62")) +
  theme_minimal()
```

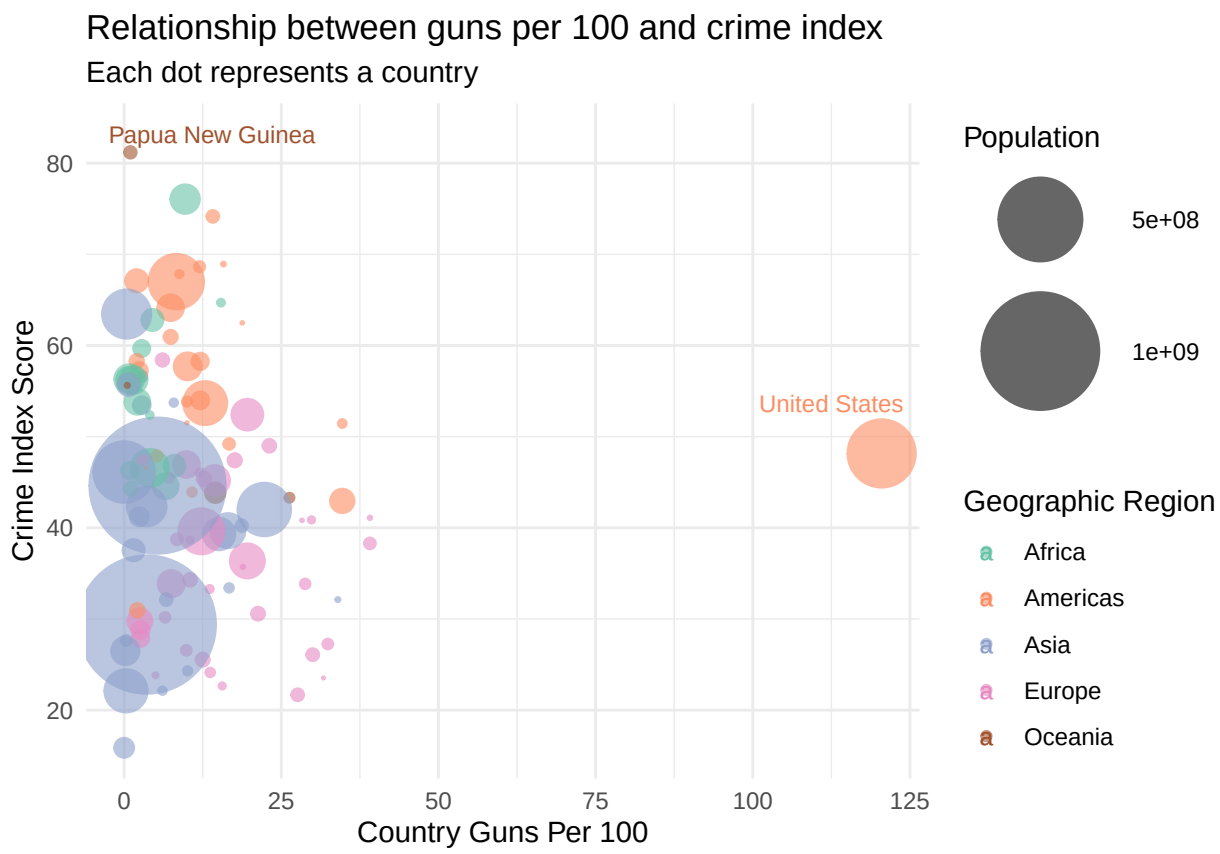



```
ggsave("plot3_regionalguns.jpg")
```

Continuing our investigation into geographic regions, the stacked bar chart shown above gives us a look into how many guns are in each region and how many of those guns are either registered or unregistered. From these three graphs, we can see that the Americas geographic region has the highest average crime index score and also is the region with the highest number of unregistered guns. Based on this, it might be easy to come to the conclusion that the number of unregistered guns explains the crime index score, but when you consider Oceania is a close second in crime index distribution, they have a very low number of total guns, and all of which look to be registered.

```
# scatterplot showing the relationship between guns per 100 and crime index

# dataframe with the outliers to make labels
outlier_gunsper100 <- selected_variables[selected_variables$guns_per_100 ==
                                         max(selected_variables$guns_per_100), ]
outlier_crimeindex <- selected_variables[selected_variables$crime_index ==
                                         max(selected_variables$crime_index), ]

# make scatterplot
ggplot(data = selected_variables,
       aes(x = guns_per_100,
           y = crime_index,
           size = population,
           color = region)) +
  geom_point(alpha = 0.6) +
```

```
scale_size(range = c(.1, 24), name = "Population") +
labs(x = "Country Guns Per 100",
     y = "Crime Index Score",
     color = "Geographic Region",
     title = "Relationship between guns per 100 and crime index",
     subtitle = "Each dot represents a country") +
scale_color_manual(values = c("Africa" = "#66c2a5",
                              "Americas" = "#fc8d62",
                              "Asia" = "#8da0cb",
                              "Europe" = "#e78ac3",
                              "Oceania" = "sienna")) +
# add text label for the guns per 100 outlier
geom_text(data = outlier_gunsper100,
          aes(label = country),
          nudge_x = -8,
          nudge_y = 5.5,
          size = 3) +
# add text for the crime index outlier
geom_text(data = outlier_crimeindex,
          aes(label = country),
          nudge_x = 13,
          nudge_y = 2,
          size = 3) +
theme_minimal()
```



```
ggsave("plot4_guns100andcrime.jpg")
```

This scatterplot shows the relationship between the number of guns per 100 people and the crime index score for all the countries. Each point on the plot represents a country, with its size indicating the country's population and its color representing the geographic region. This plot helps us identify if there is any noticeable pattern or association between the number of guns per capita and the crime index score across different countries and regions. The findings seem to agree with previous graphs in that the Americas have the most guns and are on the higher end of the crime index scale. It is also interesting that we can see all the countries of a particular region fairly grouped together in this graph.

**Regional Summary Statistics**

```
# regional gun data numerical summary (regional averages)
# and unreg to reg guns ratio for each region

# make new df
region_gun_summary <- selected_variables %>%
  # group by region
  group_by(region) %>%
  # calculate regional statistics
  summarize(avg_crime_index = mean(crime_index),
            avg_reg_guns = mean(reg_guns),
            avg_unreg_guns = mean(unreg_guns),
            # calculate gun ratio
            unreg_to_reg_ratio = sum(unreg_guns) / sum(reg_guns))

knitr::kable(region_gun_summary,
             caption = "Regional Crime and Guns Summary",
             digits = 2,
             col.names = c('Region', 'Avg Crime Index', 'Avg Reg Guns', 'Avg Unreg Guns', 'Unreg to Reg
```

Table 3: Regional Crime and Guns Summary

| Region | Avg Crime Index | Avg Reg Guns | Avg Unreg Guns | Unreg to Reg Ratio |
|---|---|---|---|---|
| Africa | 54.76 | 382760.0 | 994454.3 | 2.60 |
| Americas | 56.27 | 851906.6 | 17594173.4 | 20.65 |
| Asia | 38.40 | 1194439.5 | 7097518.8 | 5.94 |
| Europe | 36.15 | 1009174.2 | 1460501.5 | 1.45 |
| Oceania | 55.99 | 1096844.0 | 120406.0 | 0.11 |

From this table, we can see that the ratio of unregistered guns to registered guns is highest in the Americas by a long shot, followed by Asia. The ratio is lowest in Oceania. This suggests that unregistered guns may be a bigger factor in public safety in the Americas compared to other regions because the Americas have the highest crime index (region with the most crime), lowest safety index (is the least safe region), and is the region with the highest unregistered-to-registered guns ratio. However, this conclusion should be taken with caution as there may be other factors that contribute to differences.

## Statistical Analysis: Confidence Interval, Hypothesis Test, and Linear Regression Model

### Bootstrap Confidence Interval

To assess whether there is a significant difference in crime levels between countries with high and low gun ownership levels, I will be investigating the difference in crime index means between countries with a high number of guns and countries with a low number of guns. The threshold for the groups will be the median guns per 100 persons, which we found to be 9.8. Therefore, countries with greater than 9.8 guns per 100 persons will be classified as high gun countries, and countries with less than or equal to 9.8 guns per 100 persons will be classified as low gun countries.

```
# see sample of groups

knitr::kable(head(low_gun_countries),
             caption = "Low Gun Countries Group",
             col.names = c('Country', 'Guns per 100', 'Crime Idx'))
```

Table 4: Low Gun Countries Group

| | Country | Guns per 100 | Crime Idx |
|---|---|---|---|
| 89 | South Africa | 9.7 | 76.06 |
| 92 | Jamaica | 8.8 | 67.84 |
| 96 | Bulgaria | 8.4 | 38.74 |
| 97 | Brazil | 8.3 | 67.01 |
| 100 | Ghana | 8.0 | 46.81 |
| 102 | Mongolia | 7.9 | 53.73 |

```
knitr::kable(head(high_gun_countries),
             caption = "High Gun Countries Group",
             col.names = c('Country', 'Guns per 100', 'Crime Idx'))
```

Table 5: High Gun Countries Group

| | Country | Guns per 100 | Crime Idx |
|---|---|---|---|
| 1 | United States | 120.5 | 48.16 |
| 5 | Montenegro | 39.1 | 41.10 |
| 6 | Serbia | 39.1 | 38.29 |
| 7 | Canada | 34.7 | 42.95 |
| 8 | Uruguay | 34.7 | 51.44 |
| 9 | Cyprus | 34.0 | 32.12 |

```
# population difference in means
# calculate the mean crime_index for each group
# and calculate the observed difference in means between the two groups

# show favstats of crime index for low gun countries - mean = 46.95538
#knitr::kable(favstats(~crime_index, data=low_gun_countries),
#             caption = "Low Gun Countries Group Summary",
#             digits = 1)
```

```
# show favstats of crime index for high gun countries - mean = 42.59788
#knitr::kable(favstats(~crime_index, data=high_gun_countries),
#             caption = "High Gun Countries Group Summary",
#             digits = 1)

low_group_summary = favstats(~crime_index, data = low_gun_countries)
high_group_summary = favstats(~crime_index, data = high_gun_countries)


# calculate mean crime index of low gun countries
low_gun_mean_crime <- mean(~crime_index, data=low_gun_countries)
# calculate mean crime index of high gun countries
high_gun_mean_crime <- mean(~crime_index, data=high_gun_countries)

# calculate observed difference in crime index for low and high gun countries
diff_in_means <- high_gun_mean_crime - low_gun_mean_crime
#diff_in_means # -4.3575

# combine group summaries to print table
group_summaries_combined = rbind(low_group_summary, high_group_summary)
# add row names
rownames(group_summaries_combined) = c("Low Gun Countries", "High Gun Countries")
knitr::kable(group_summaries_combined,
             caption = "Crime Index Summary for Both Groups",
             digits = 2,
             col.names = c('Min', 'Q1', 'Median', 'Q3',
                           'Max', 'Mean', 'SD', 'n', 'NAs'))
```

Table 6: Crime Index Summary for Both Groups

| | Min | Q1 | Median | Q3 | Max | Mean | SD | n | NAs |
|---|---|---|---|---|---|---|---|---|---|
| Low Gun Countries | 15.87 | 33.43 | 46.69 | 56.75 | 81.19 | 46.96 | 14.99 | 52 | 0 |
| High Gun Countries | 21.68 | 33.74 | 41.56 | 49.76 | 74.16 | 42.60 | 12.69 | 52 | 0 |

We found the observed difference in mean crime index between the low gun countries and high gun countries groups was -4.3575 by subtracting the mean crime index from the low gun countries group from the mean crime index of the high gun countries group. The result shows that the crime index for the low gun countries group is higher (less safe) than the high gun countries group.

```
# bootstrapping to estimate the sampling distribution of crime index for groups

# bootstrapping for low gun countries
bootstrap_samplemeans_low =
  do(500)*mean(~crime_index,
               data = sample_n(low_gun_countries,
                               size = 52,
                               replace = TRUE))
bootstrap_samplemeans_low = bootstrap_samplemeans_low %>%
  rename(bootstrap_samplemean_lowguns = mean)

# bootstrapping for high gun countries
bootstrap_samplemeans_high =
```

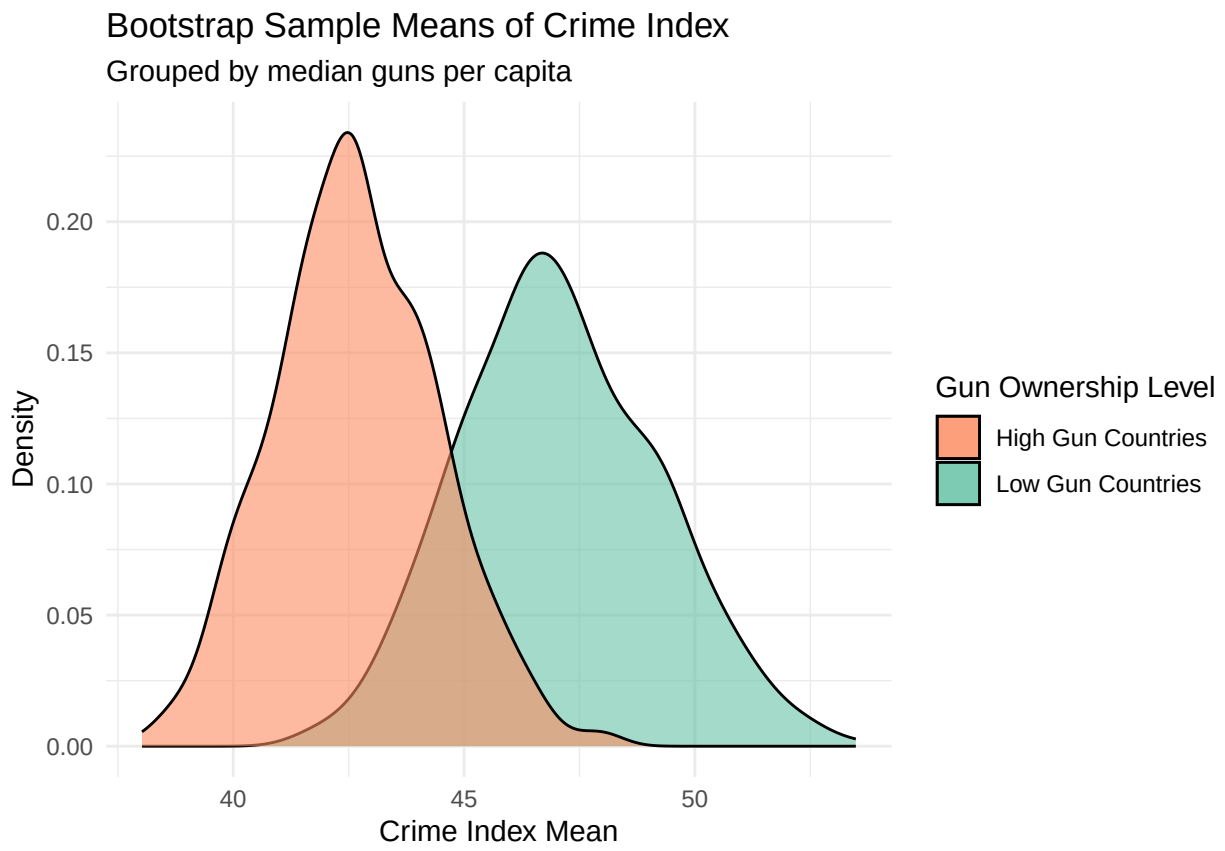
```
  do(500)*mean(~crime_index,
               data = sample_n(high_gun_countries,
                               size = 52,
                               replace = TRUE))
bootstrap_samplemeans_high = bootstrap_samplemeans_high %>%
  rename(bootstrap_samplemean_highguns = mean)

# combine the two dataframes into one
bootstrap_samplemeans_combined = cbind(bootstrap_samplemeans_low, bootstrap_samplemeans_high)

# print bootstrap sample means of both groups
knitr::kable(head(bootstrap_samplemeans_combined),
             caption = 'Bootstrapped Sample Means for Both Groups',
             digits = 2,
             col.names = c('Low Guns Group', 'High Guns Group'))
```

Table 7: Bootstrapped Sample Means for Both Groups

| Low Guns Group | High Guns Group |
|---|---|
| 45.09 | 40.17 |
| 51.55 | 41.95 |
| 45.41 | 42.31 |
| 47.25 | 41.25 |
| 47.11 | 43.95 |
| 46.38 | 44.43 |

Here, we can see a sample of what the bootstrapped sample means were for both groups. Now, let's get a better look at them in the following density plot.

```
# density plot of results from bootstrapped sample means for groups

ggplot() +
  # low gun countries
  geom_density(data = bootstrap_samplemeans_low,
               aes(x = bootstrap_samplemean_lowguns,
                   fill = "Low Gun Countries"),
               alpha = 0.6) +
  # high gun countries
  geom_density(data = bootstrap_samplemeans_high,
               aes(x = bootstrap_samplemean_highguns,
                   fill = "High Gun Countries"),
               alpha = 0.6) +
  labs(
    title = "Bootstrap Sample Means of Crime Index",
    subtitle = "Grouped by median guns per capita",
    x = "Crime Index Mean",
    y = "Density",
    fill = "Gun Ownership Level"
  ) +
  scale_fill_manual(values = c("Low Gun Countries" = "#66c2a5",
                               "High Gun Countries" = "#fc8d62")) +
  theme_minimal()
```




```
ggsave("plot5_bootstrap.jpg")
```

What we can see from the density plot of bootstrapped crime index sample means is that the density for crime index score for low gun countries is higher than that of high gun countries, but there is a significant amount of overlap. This overlap may suggest that there is not a significant difference in crime index scores between countries with a low number of guns and countries with a high number of guns.

```
# find confidence interval of bootstrapped crime index means for groups

# calculate CI for low gun country group
confidence_int_low = quantile(bootstrap_samplemeans_low$bootstrap_samplemean_lowguns,
                              probs = c(.025, .975))
# calculate CI for high gun country group
confidence_int_high = quantile(bootstrap_samplemeans_high$bootstrap_samplemean_highguns,
                              probs = c(.025, .975))

# combine CI's into one df
conf_ints <- data.frame(confidence_int_low,
                        confidence_int_high)

# rename columns
colnames(conf_ints) <- c("Low Guns",
                         "High Guns")

# print the CI results in a table
```

```
knitr::kable(conf_ints,
             caption = "95% CI for Crime Index in Low and High Guns Groups",
             digits = 2)
```

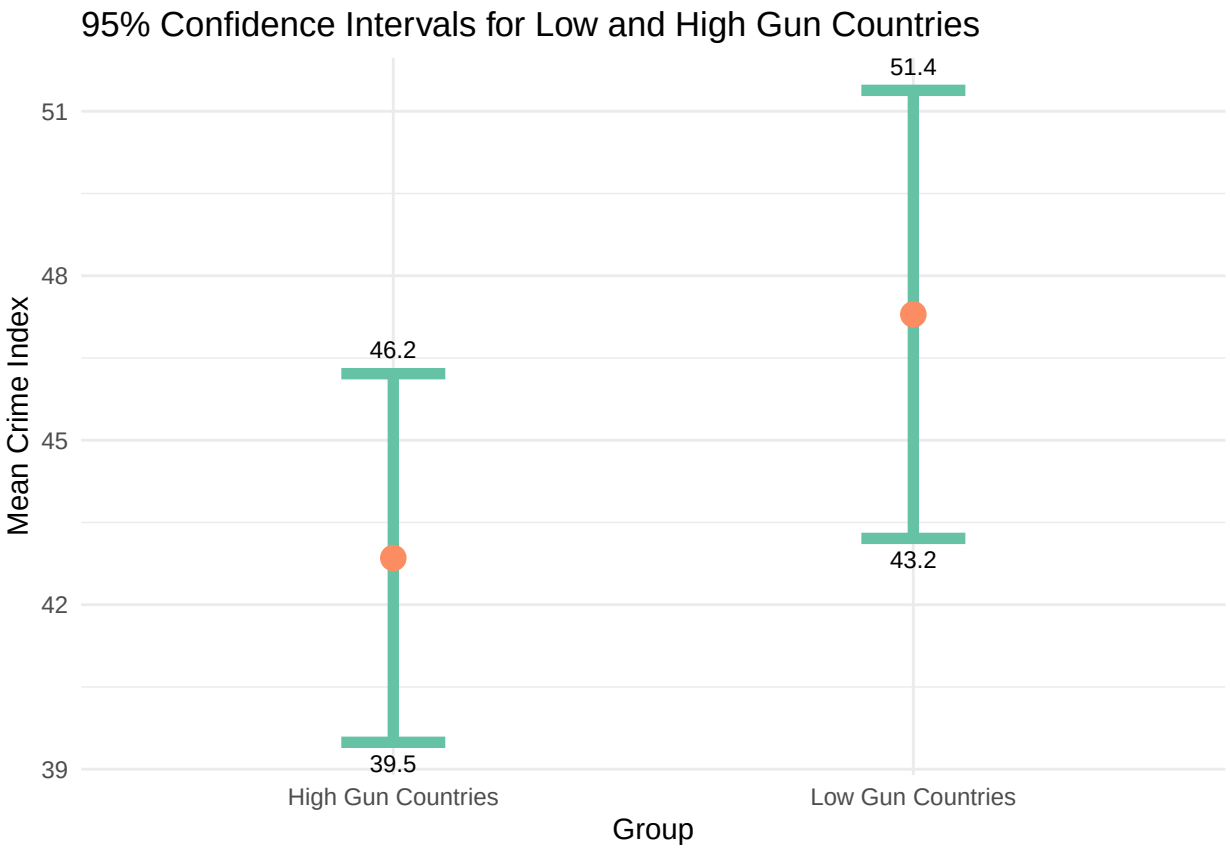
Table 8: 95% CI for Crime Index in Low and High Guns Groups

| | Low Guns | High Guns |
|---|---|---|
| 2.5% | 43.21 | 39.49 |
| 97.5% | 51.38 | 46.21 |

Based on the calculated confidence intervals, we are 95% confident that the true population is between 43.23863 and 50.67331 of our confidence interval for the countries with a low number of guns, and between 39.45961 and 45.86645 for the countries with a high number of guns. To get a better look at the difference in confidence intervals between the groups, we can look at the following interval plot.

```
# interval plot for confidence intervals

# calculate mean values
mean_low <- mean(confidence_int_low)
mean_high <- mean(confidence_int_high)

# combine CI data into a dataframe
ci_data_combined <- data.frame(
  group = c("Low Gun Countries", "High Gun Countries"),
  mean_value = c(mean_low, mean_high),
  lower_bound = c(confidence_int_low[1], confidence_int_high[1]),
  upper_bound = c(confidence_int_low[2], confidence_int_high[2])
)

# make interval plot to show overlap in CI's
ggplot(ci_data_combined,
       aes(x = group,
           y = lower_bound)) +
  # line chart to connect the two CI means
  #geom_line(aes(x = group,
   #              y = mean_value),
    #        group = 1) +
  # plot confidence interval bars
  geom_errorbar(aes(ymin = lower_bound,
                    ymax = upper_bound),
                width = 0.2,
                color = "#66c2a5",
                size = 2) +
  # point for mean crime index for each group
  geom_point(aes(y = mean_value),
             color = "#fc8d62",
             size = 4) +
  # add text to indicate upper/lower bounds
  geom_text(aes(y = lower_bound,
                label = round(lower_bound, 1)),
            vjust = 1.7,
```

```
            size = 3) +
  geom_text(aes(y = upper_bound,
                label = round(upper_bound, 1)),
            vjust = -0.9,
            size = 3) +
  labs(title = "95% Confidence Intervals for Low and High Gun Countries",
       x = "Group",
       y = "Mean Crime Index") +
  theme_minimal()
```




```
ggsave("plot6_ci.jpg")
```

The plot above shows the 95% confidence intervals for the two groups. In this visualization, we can easily see there is a notable amount of overlap between the two confidence intervals, suggesting that there is no evidence that there is a significant difference in crime index scores between countries with a high number of guns and countries with low number of guns. To test this conclusion, we performed the following hypothesis test.

### Hypothesis Test

The population is a hypothetical population of all hypothetical countries in 2017. Let $\mu_1$ be the mean crime index score for low gun countries and let $\mu_2$ be the mean crime index score for high gun countries.

- $H_0 : \mu_1 = \mu_2$

- $H_A : \mu_1 \neq \mu_2$
- $H_0$ : The mean crime index of all low gun countries is equal to the mean crime index of all high gun countries.
- $H_A$ : The mean crime index of all low gun countries is not equal to the mean crime index of all high gun countries.

```
# remove country and guns_per_100 from groups to get vectors with crime index

crime_idx_low_gun_countries = subset(low_gun_countries, select = -c(country,guns_per_100))
crime_idx_high_gun_countries = subset(high_gun_countries, select = -c(country,guns_per_100))
```

```
# perform two sample t-test

results <- t.test(crime_idx_low_gun_countries$crime_index,
                  crime_idx_high_gun_countries$crime_index)
#results

# make a data frame to show better results using knitr when i knit to pdf
t_test_results <- data.frame(
  t_value = round(results$statistic, 4),
  df = round(results$parameter, 4),
  p_value = round(results$p.value, 4),
  conf_low = round(results$conf.int[1], 4),
  conf_high = round(results$conf.int[2], 4)
)

# rename columns
colnames(t_test_results) <- c('T-value',
                              'DF',
                              'P-value',
                              'Conf Low',
                              'Conf High')

knitr::kable(t_test_results, row.names = FALSE,
             caption = "Two-Sample T-Test Results",
             digits = 3)
```

Table 9: Two-Sample T-Test Results

| T-value | DF | P-value | Conf Low | Conf High |
|---|---|---|---|---|
| 1.6 | 99.295 | 0.113 | -1.046 | 9.761 |

We fail to reject the null hypothesis that the mean crime index for all low gun countries is equal to the mean crime index for all high gun countries. In other words, the data does not provide convincing evidence at the 0.05 significance level that the mean crime index for low gun countries is not equal to the mean crime index for high gun countries.

### Linear Regression

The following regression model aims to assess the linear relationship between the number of unregistered guns and crime index of countries in different regions. The response variable (dependent variable) is `crime_index`,

while the predictor variables (independent variables) are `region` (a categorical variable) and the number of unregistered guns, `unreg_guns` (a continuous variable). The model is fitted using linear regression, and the summary of the model is printed to provide insights into the model's performance and significance of variables.

Crime index is the dependent variable that we are trying to predict. $\beta_0$ is the intercept term, and $\beta_1$ and $\beta_2$ are the coefficients for the region and unregistered guns variables, respectively. The $\varepsilon$ represents the error in the prediction. This gives us the following model formula:

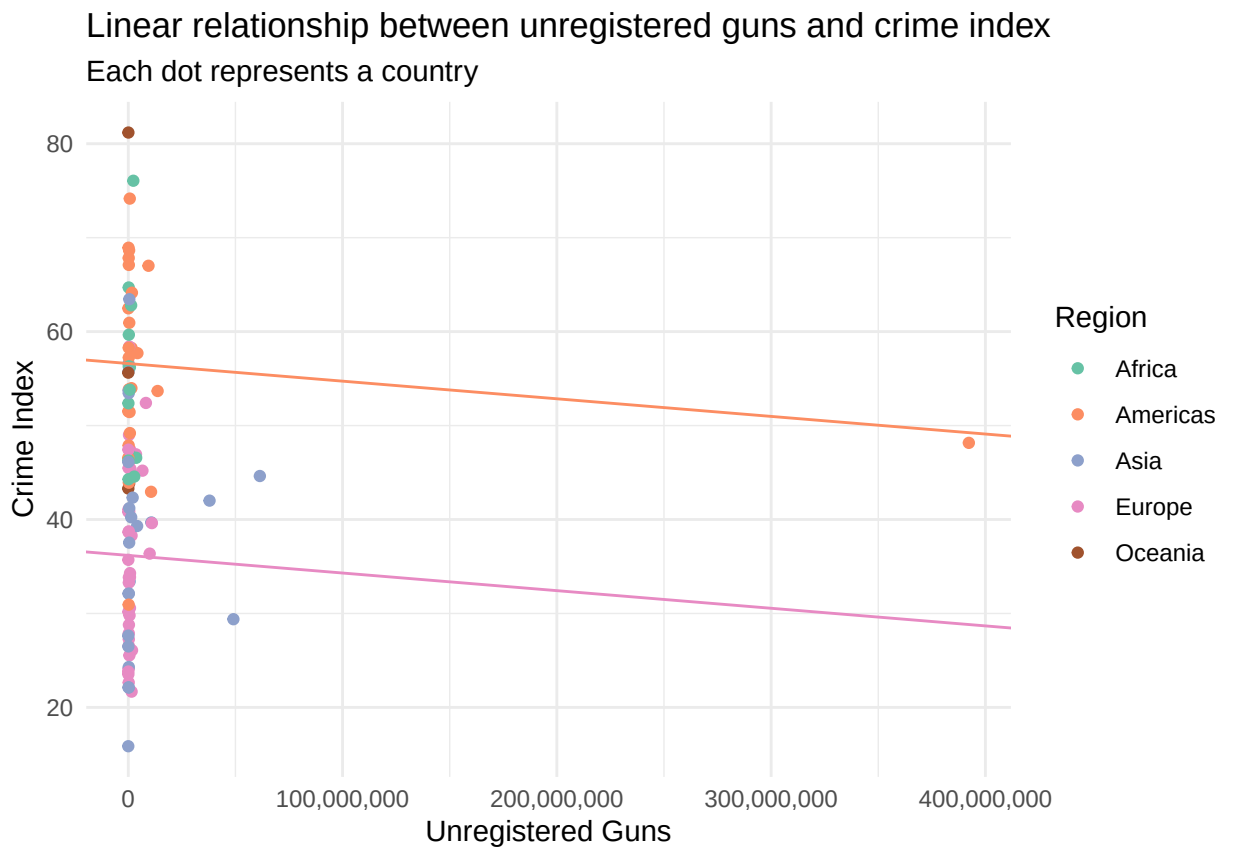
$$\texttt{crime\_index} = \beta_0 + \beta_1 * \texttt{region} + \beta_2 * \texttt{unreg\_guns} + \varepsilon$$

```
# linear regression model fitting
# response variable (y) is crime_index
# predictor variables (x) is region and unreg_guns

# fit the model
fittedmodel <- lm(crime_index ~ region + unreg_guns, data = selected_variables)

# print model summary
summary(fittedmodel)
```

```
##
## Call:
## lm(formula = crime_index ~ region + unreg_guns, data = selected_variables)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.6604  -8.2412   0.0061   7.5730  25.1962
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.478e+01  2.855e+00  19.191  < 2e-16 ***
## regionAmericas  1.821e+00  3.594e+00   0.507    0.613
## regionAsia     -1.625e+01  3.596e+00  -4.519 1.74e-05 ***
## regionEurope   -1.860e+01  3.351e+00  -5.550 2.43e-07 ***
## regionOceania   1.213e+00  6.055e+00   0.200    0.842
## unreg_guns     -1.876e-08  2.727e-08  -0.688    0.493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.68 on 98 degrees of freedom
## Multiple R-squared:  0.4456, Adjusted R-squared:  0.4173
## F-statistic: 15.75 on 5 and 98 DF,  p-value: 2.318e-11
```

From the summary, we find the F-statistic is 15.75 and the p-value associated with it is 2.318e-11, which is very low. This indicates that the overall model is statistically significant, and at least one of the predictor variable has a significant effect on crime index. We also found that the number of unregistered guns does not have a significant effect on crime index when considering other variables in the model, because it's p-value is greater than 0.05, and that the only significant regions are Asia and Europe. This means that, since Africa was our reference, the crime index for a country in Asia would be significantly lower by about 16.25 points than a country in Africa, and the crime index for a country in Europe would be significantly lower by about 18.60 points than a country in Africa when controlling for the number of unregistered guns.

```r
# plotting the fitted model with regression lines (for Americas and Europe regions)
# the plot compares the most dangerous and least dangerous regions, Americas and Europe

ggplot(data = selected_variables,
       aes(x = unreg_guns,
           y = crime_index,
           color = region)) +
  geom_point() +
  # regression line for Americas region
  geom_abline(intercept = fittedmodel$coefficients[1] + fittedmodel$coefficients[2],
              slope = fittedmodel$coefficients[6],
              color = '#fc8d62') +
  # regression line for Europe region
  geom_abline(intercept = fittedmodel$coefficients[1] + fittedmodel$coefficients[4],
              slope = fittedmodel$coefficients[6],
              color = '#e78ac3') +
  # show numbers instead of scientific notation
  scale_x_continuous(labels = scales::number_format(big.mark = ",")) +
  scale_color_manual(values = c("Africa" = "#66c2a5",
                                "Americas" = "#fc8d62",
                                "Asia" = "#8da0cb",
                                "Europe" = "#e78ac3",
                                "Oceania" = "sienna")) +
  labs(title = "Linear relationship between unregistered guns and crime index",
       subtitle = "Each dot represents a country",
       x = "Unregistered Guns",
       y = "Crime Index",
       color = "Region") +
  theme_minimal()
```




```
ggsave("plot7_linreg1.jpg")
```

The above plot visualizes the linear relationship between the number of unregistered guns and crime index scores for countries in the Americas region and Europe region, the most dangerous and least dangerous regions. The plot suggests that there is a negative relationship between the number of unregistered guns and crime index scores within those regions, meaning that as the number of unregistered guns increases, the crime index scores tend to decrease, according to the fitted regression lines. However, we can see there is an extreme horizontal outlier, which is altering the scale of the graph and may be causing biased estimates of the coefficients and affecting the model's overall performance, so a conclusion should not be made because the fitted model is not appropriate by looking at the graph since the extreme horizontal outlier in the graph is a high leverage point.

**Removing outlier**

Since we found an extreme outlier in the regression analysis plot, I will be performing a sensitivity analysis by running the same regression model without the outlier country (United States) and assessing the influence of the outlier on the coefficients, the model's overall performance, and the plot.

```
# linear regression model fitting with outlier removed
# response variable is crime_index
# predictor variables are region and unreg_guns

# fit the model
fittedmodel2 <- lm(crime_index ~ region + unreg_guns, data = remove_us_outlier)
```

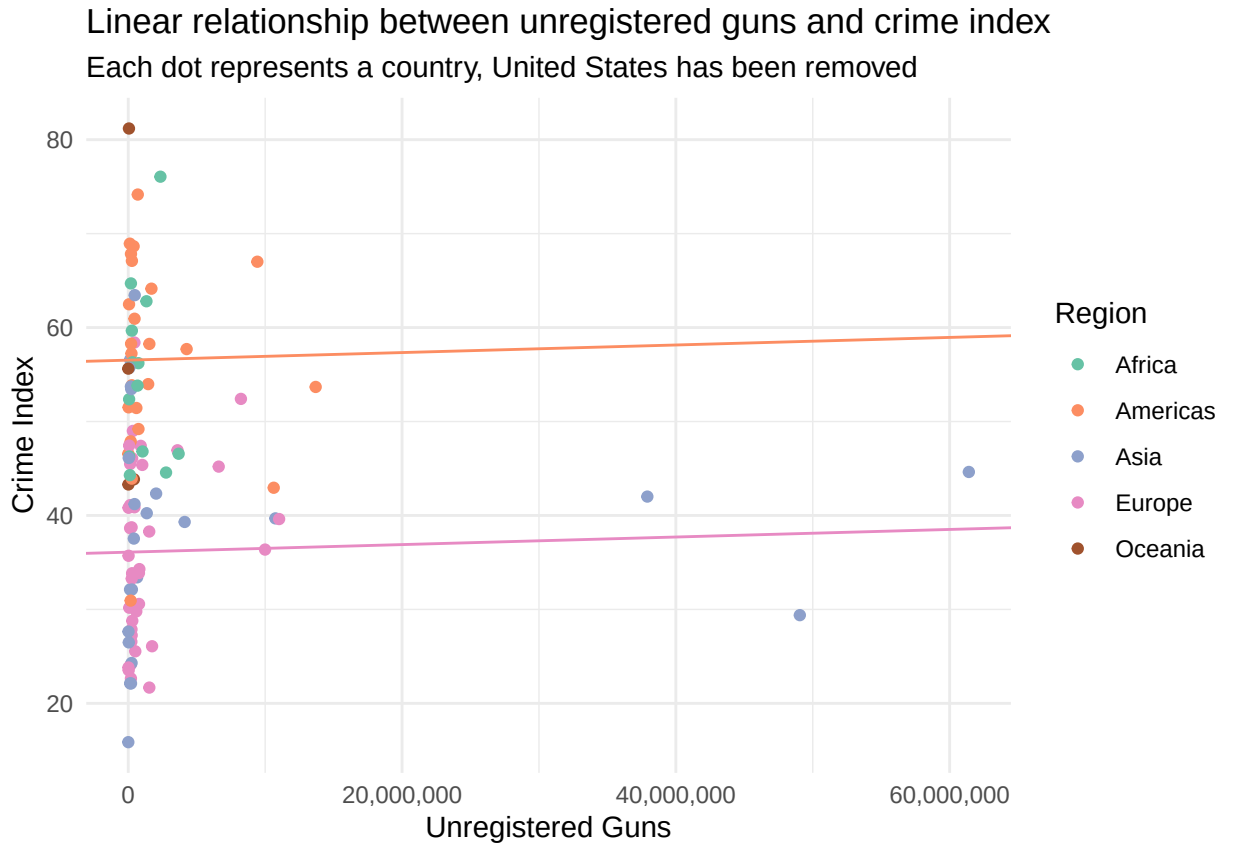
```r
# print model summary
summary(fittedmodel2)
```

```
##
## Call:
## lm(formula = crime_index ~ region + unreg_guns, data = remove_us_outlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5988  -8.3741  -0.1271   7.7722  25.3170
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.472e+01  2.869e+00  19.077  < 2e-16 ***
## regionAmericas  1.808e+00  3.608e+00   0.501    0.617
## regionAsia     -1.661e+01  3.688e+00  -4.504 1.86e-05 ***
## regionEurope   -1.863e+01  3.365e+00  -5.536 2.64e-07 ***
## regionOceania   1.264e+00  6.080e+00   0.208    0.836
## unreg_guns      4.031e-08  1.267e-07   0.318    0.751
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.72 on 97 degrees of freedom
## Multiple R-squared:  0.4466, Adjusted R-squared:  0.4181
## F-statistic: 15.65 on 5 and 97 DF,  p-value: 2.824e-11
```

From the updated summary after removing the United States, we found that some of the results were very similar, but the estimate of `unreg_guns` actually flips signs, but is still very close to zero. We again find that the overall model is statistically significant, and at least one of the predictor variables has a significant effect on crime index. Although the coefficient for unregistered guns is greater than the previous model, the p-value increase from 0.493 to 0.751, which means we still find that the number of unregistered guns does not have a significant effect on crime index. We also still find that the only significant regions are Asia and Europe.

```r
# plot the regression model without the outlier country

# plotting the fitted model with regression lines (for Americas and Europe regions)
# the plot compares the most dangerous and least dangerous regions, Americas and Europe

ggplot(data = remove_us_outlier,
       aes(x = unreg_guns,
           y = crime_index,
           color = region)) +
  geom_point() +
  # regression line for Americas region
  geom_abline(intercept = fittedmodel2$coefficients[1] + fittedmodel2$coefficients[2],
              slope = fittedmodel2$coefficients[6],
              color = '#fc8d62') +
  # regression line for Europe region
  geom_abline(intercept = fittedmodel2$coefficients[1] + fittedmodel2$coefficients[4],
              slope = fittedmodel2$coefficients[6],
              color = '#e78ac3') +
  # show numbers instead of scientific notation
```

```
  scale_x_continuous(labels = scales::number_format(big.mark = ",")) +
  scale_color_manual(values = c("Africa" = "#66c2a5",
                                "Americas" = "#fc8d62",
                                "Asia" = "#8da0cb",
                                "Europe" = "#e78ac3",
                                "Oceania" = "sienna")) +
  labs(title = "Linear relationship between unregistered guns and crime index",
       subtitle = "Each dot represents a country, United States has been removed",
       x = "Unregistered Guns",
       y = "Crime Index",
       color = "Region") +
  theme_minimal()
```




```
ggsave("plot8_linreg2.jpg")
```

After removing the extreme horizontal outlier, the United States, the plot showing the linear relationship between the number of unregistered guns and crime index scored for countries in the Americas and Europe regions suggests that there is now a slightly positive relationship between those variables within those regions. This means that, contrary to the original plot, as the number of unregistered guns increases, the crime index scores tend to increase, according to the fitted regression lines.

## Regression Model Evaluation

To understand the performance of the two linear models, we then compared the two in terms of their mean square error (MSE). To do this, we split the data into training and test sets, trained both models on the

same training set, calculated the mean square error for both models on the test set, and analyzed the results.

```
# compare the two models and calculate MSE

total = nrow(selected_variables)
total2 = nrow(remove_us_outlier)
n = round(.7*total)
n2 = round(.7*total2)

set.seed(101)

# split data (70/30 train/test split)
split_index = sample(1:total, size = n)
split_index2 = sample(1:total2, size = n2)
train = selected_variables[split_index, ]
test = selected_variables[-split_index, ]
train2 = remove_us_outlier[split_index2, ]
test2 = remove_us_outlier[-split_index2, ]

# fit both models using training data
fittedmodel = lm(crime_index ~ region + unreg_guns, data = train)
fittedmodel2 = lm(crime_index ~ region + unreg_guns, data = train2)

# predict crime index for test data using both models
pred = predict(fittedmodel, newdata = test)
pred2 = predict(fittedmodel2, newdata = test2)

# calculate MSE for both models
mse1 = mean((test$crime_index - pred)^2)
mse2 = mean((test2$crime_index - pred2)^2)

# store MSE results
mse_values <- c(mse1, mse2)
names(mse_values) <- c("MSE for fittedmodel (with outlier)",
                       "MSE for fittedmodel2 (without outlier)")

# print results using knitr
knitr::kable(mse_values,
             caption = "MSE Results",
             digits = 1)
```

Table 10: MSE Results

| | x |
|---|---|
| MSE for fittedmodel (with outlier) | 133.2 |
| MSE for fittedmodel2 (without outlier) | 98.0 |

The results show that `fittedmodel` (with the outlier) yielded an MSE of 133.2, while `fittedmodel2` (without the outlier) had an MSE of 98.0. Although the `fittedmodel` demonstrated a marginally lower MSE compared to `fittedmodel2`, indicating a significant improvement in predictive performance. Based on this, we can conclude that the removal of the United States as an outlier did improve the model.

## Discussion and Conclusion

Throughout the project, I looked at how the crime and safety index of countries, the amount of registered and unregistered guns, and the region and subregion they belonged to may affect one another. There are some trends that can be seen within the visualized data, such as higher safety index regions tending to have a larger proportion of guns per civilian than those with a lower safety index.

From the bootstrap confidence interval, the data seems to suggest that countries with a low amount of guns may have a higher crime index than those with a high amount of guns, though much of the difference could be due to random chance. The difference in crime index being random is also supported by the hypothesis test posted above, in which there was insufficient evidence to reject the null hypothesis that the mean crime index is equal between low gun countries and high gun countries. The regression line showing the linear relationship between registered guns and crime index shows a somewhat weak positive correlation between the two for the Americas and European regions.

Both within individual countries and whole regions, however, there is no noticeable trend between ownership of guns and crime index; the top 10 countries with a high crime index did not have a consistently high nor low amount of guns, registered or unregistered. This is most evident with the top two, Papua New Guinea and South Africa, the latter of which has almost 10 times the amount of guns per citizen and about 110 times the amount of guns total, yet is relatively close in crime index. As well, the regions with the highest and the lowest crime indexes had varying amounts of guns per 100 people. This suggests that if there is a relationship between guns and crime index, it may not be linear. It should be worth noting that some countries' information was removed due to a lack of complete data; if their information was included with accurate numbers, the pattern we see may differ significantly.