

eda

Eleanor Jiang 305002785

2024-11-17

EDA

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

##           name           age           gender
##           0             0             0
##   raceethnicity         month           day
##           0             0             0
##           year      streetaddress       city
##           0             1             0
##           state      latitude      longitude
##           0             0             0
##           state_fp      county_fp      tract_ce
##           0             0             0
##           geo_id      county_id      namelsad
##           0             0             0
## lawenforcementagency      cause      armed
##           0             0             0
##           pop      share_white      share_black
##           0             0             0
##           share_hispanic      p_income      h_income
##           0             0             0
##           county_income      comp_income      county_bucket
##           0             0             30
##           nat_bucket      pov      urate
##           0             0             0
##           college
##           0

## Warning: NAs introduced by coercion
## [1] "numeric"
## [1] 4

## Warning: NAs introduced by coercion

The indicator age contains missing values "Unknown". We choose to fill them with the average value.
```

```
glimpse(new_data)
```

```
## Rows: 503
## Columns: 35
## $ name      <chr> "Elton Simpson", "William Chapman II", "James Co~
## $ age       <dbl> 30, 18, 43, 50, 38, 35, 17, 24, 59, 32, 22, 46, 2~
## $ gender    <chr> "Male", "Male", "Male", "Male", "Male", "Male", "~
## $ raceethnicity <chr> "Black", "Black", "White", "Black", "Black", "Whi~
## $ month     <chr> "May", "April", "May", "May", "February", "Februa~
## $ day       <int> 3, 22, 20, 31, 20, 13, 22, 13, 21, 12, 9, 12, 25,~
## $ year      <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2~
## $ streetaddress <chr> "4999 Naaman Forest Blvd", "1098 Frederick Blvd",~
## $ city      <chr> "Garland", "Portsmouth", "Charleston", "Rochester~
## $ state     <chr> "TX", "VA", "SC", "NY", "AL", "PA", "TX", "TX", "~
## $ latitude  <dbl> 32.95994, 36.82901, 32.85778, 43.14785, 33.48143,~
## $ longitude <dbl> -96.63896, -76.34144, -80.07766, -77.63095, -86.8~
## $ state_fp  <int> 48, 51, 45, 36, 1, 42, 48, 48, 45, 6, 12, 15, 48,~
## $ county_fp <int> 113, 740, 19, 55, 73, 11, 183, 201, 63, 37, 113, ~
## $ tract_ce  <int> 19027, 211500, 5700, 2700, 5000, 2000, 1100, 2401~
## $ geo_id    <dbl> 48113019027, 51740211500, 45019005700, 3605500270~
## $ county_id <int> 48113, 51740, 45019, 36055, 1073, 42011, 48183, 4~
## $ namelsad  <chr> "Census Tract 190.27", "Census Tract 2115", "Cens~
## $ lawenforcementagency <chr> "Garland Police Department", "Portsmouth Police D~
## $ cause     <chr> "Gunshot", "Gunshot", "Gunshot", "Taser", "Gunsho~
## $ armed     <chr> "No", "No", "Knife", "Firearm", "Firearm", "Firea~
## $ pop       <int> 4775, 1640, 8668, 1271, 3681, 4017, 4045, 4049, 3~
## $ share_white <dbl> 34.7, 40.9, 85.5, 0.6, 44.4, 37.4, 44.8, 6.5, 71.~
## $ share_black <dbl> 16.3, 53.8, 11.0, 95.6, 22.4, 10.7, 34.1, 31.8, 1~
## $ share_hispanic <dbl> 14.6, 0.0, 0.7, 3.9, 28.9, 47.7, 19.7, 58.6, 2.7,~
## $ p_income  <int> 31009, 25262, 38810, 11558, 21908, 20761, 14332, ~
## $ h_income  <int> 49973, 27418, 80891, 18833, 35780, 29707, 26458, ~
## $ county_income <int> 49481, 46166, 50792, 52394, 45429, 55170, 45525, ~
## $ comp_income <dbl> 1.0099432, 0.5939003, 1.5925933, 0.3594496, 0.787~
## $ county_bucket <int> 3, 1, 5, NA, 2, 1, 1, 1, 5, 1, 3, 2, 5, 2, 5, 3, ~
## $ nat_bucket <int> 3, 1, 5, 1, 2, 1, 1, 1, 5, 1, 4, 4, 4, 2, 4, 3, 1~
## $ pov       <dbl> 9.2, 35.2, 4.0, 49.9, 23.2, 36.6, 27.4, 40.9, 14.~
## $ urate     <dbl> 0.09214891, 0.15204678, 0.09204239, 0.25925926, 0~
## $ college   <dbl> 0.31563891, 0.12055336, 0.49587195, 0.09653092, 0~
## $ age_num   <dbl> 30, 18, 43, 50, 38, 35, 17, 24, 59, 32, 22, 46, 2~
```

```
##### Research question 1
```

```
## 1. Race/Ethnicity
```

```
race_counts <- table(new_data$raceethnicity)
race_percentages <- prop.table(race_counts) * 100
```

```
cat("Race/Ethnicity Distribution:\n")
```

```
## Race/Ethnicity Distribution:
```

```
print(race_counts)
```

```
##
## Asian/Pacific Islander      Black      Hispanic/Latino
##                5              156              73
##      Native American      Unknown              White
```

```
##                                     5                                     13                                     251
cat("\nPercentages:\n")

##
## Percentages:
print(race_percentages)

##
## Asian/Pacific Islander          Black          Hispanic/Latino
##          0.9940358          31.0139165          14.5129225
##          Native American          Unknown          White
##          0.9940358          2.5844930          49.9005964

plot1 <- ggplot(new_data, aes(x = raceethnicity)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of Race/Ethnicity", x = "Race/Ethnicity", y = "Count")

# Plot 1: Relationship between 'race' and 'poverty rate'
plot2 <- ggplot(new_data, aes(x = raceethnicity, y = pov)) +
  geom_boxplot(fill = "lightgreen", outlier.colour = "red", outlier.shape = 1) +
  geom_jitter(width = 0.2, alpha = 0.2) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Relationship between Race/Ethnicity and Poverty Rate",
        x = "Race/Ethnicity",
        y = "Poverty Rate")

#### Research question 2

## 6. Armed Status
armed_counts <- table(new_data$armed)
armed_percentages <- prop.table(armed_counts) * 100

cat("\nArmed Status Distribution:\n")

##
## Armed Status Distribution:
print(armed_counts)

##
##          Disputed          Firearm          Knife          No
##          1          269          67          104
## Non-lethal firearm          Other          Unknown          Vehicle
##          16          30          5          11

cat("\nPercentages:\n")

##
## Percentages:
print(armed_percentages)

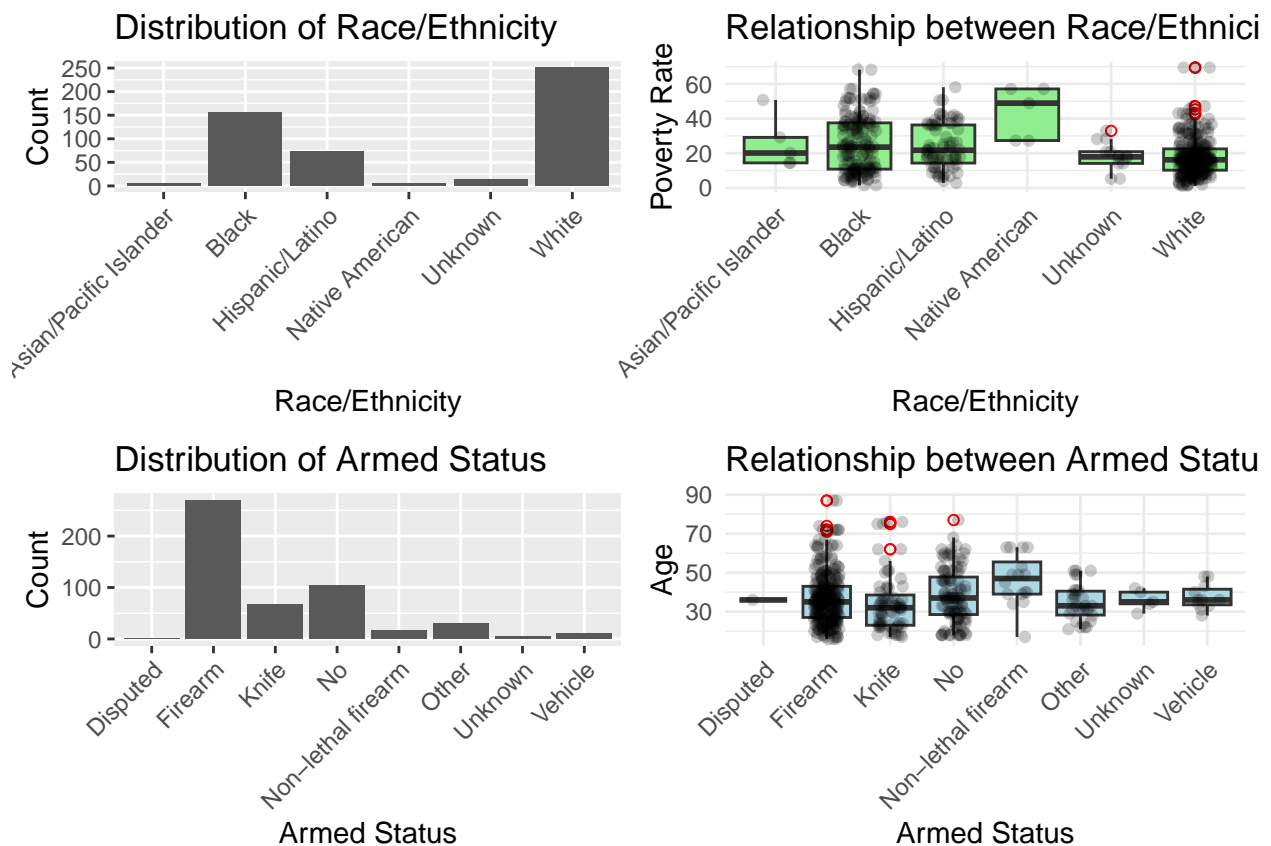
##
##          Disputed          Firearm          Knife          No
```

```
##          0.1988072          53.4791252          13.3200795          20.6759443
## Non-lethal firearm          Other          Unknown          Vehicle
##          3.1809145          5.9642147          0.9940358          2.1868787
```

```
plot3 <- ggplot(new_data, aes(x = armed)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of Armed Status", x = "Armed Status", y = "Count")
```

```
# Plot 2: Relationship between 'armed' and 'age'
plot4 <- ggplot(new_data, aes(x = armed, y = age)) +
  geom_boxplot(fill = "lightblue", outlier.colour = "red", outlier.shape = 1) +
  geom_jitter(width = 0.2, alpha = 0.2) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Relationship between Armed Status and Age",
       x = "Armed Status",
       y = "Age")
```

```
library(cowplot)
plot_grid(plot1, plot2, plot3, plot4, ncol=2)
```



There are mostly black and white in race/ethnicity in police killings. We can see that the average poverty rate is lower in white compared to black. Moreover, we can observe that there are many people armed with firearm and the range of the age of the deceased is wide from less than 30 to 90.