Murder Gun

Ladan Foroughi

10/05/2021

All the library were used for analyzing of this dataset as follows:

```
if(!require(pacman))install.packages("pacman")
pacman::p_load(
  tidyverse,
  dplyr,
  ggplot,
  caret,
  magnittr,
  pacman,
  GGally,
  knitr,
  parallel,
  rattel,
  tictoc,
  gridExtra,
  kableExtra,
  readr,
  purrr,
  randomForest,
  pROC,
  fastDummies,
  rpart.plot,
  data.table,
  reshape2,
  graphics,
  corrplot,
  latexpdf,
  ReporteRs,
  tinytex,
  latexdiffr,
  latex2exp,
  ggrepel,
)
```

The US murders dataset is included as part of the “dslabs” package.

```
filename <- "murders.csv"
dir <- system.file("extdata", package = "dslabs")
fullpath <- file.path(dir, filename)
#file.copy(fullpath, "murders.csv")
murder <- read_csv(filename)
```

This dataset is consist of 5 column and 51 rows (Table 1). The name of columns are: State, Abbreviation(Abb), Region, Population and total (as shown in Table 2). There is no missing data in dataset (Table 3). The six first row of dataset is shown in Table 4:

Table 1: The dimention of murder dataset

x
51
5

Table 2: The name of columns in murder dataset

x
state
abb
region
population
total

Table 3: The number of missing data in murder dataset

x
FALSE

Table 4: The first six rows of murder dataset

state	abb	region	population	total
Alabama	AL	South	4779736	135
Alaska	AK	West	710231	19
Arizona	AZ	West	6392017	232
Arkansas	AR	South	2915918	93
California	CA	West	37253956	1257
Colorado	CO	West	5029196	65

The first six rows of state names by order of their total murders as shown in Table 5.

```
ind <- order(murder$total)
State_order_murder <- murder$state[ind]

kable(head(State_order_murder),
      "pandoc",
      caption = "The State names based on order of Their total          murder",
      align = "c",
      font_size = 5)
```

Table 5: The State names based on order of Their total murder

x
Vermont
North Dakota
New Hampshire
Wyoming
Hawaii
South Dakota

It is shown that the maximum total murder is 1257 in California with population around 37253956 and minimum total murder is 2 in Vermont with population around 625741 (Table 6).

```r
# the max and min of total murder
max_total_murder <- max(murder$total)
min_total_murder <- min(murder$total)

# The state name of min and max of total number
i_max <- which.max(murder$total)
State_max_total_murdre <- murder$state[i_max]
i_min <- which.min(murder$total)
state_min_total_murdre <- murder$state[i_min]

# the population of max and min of total number state
Population_max_total_murder <- murder$population[i_max]
Population_min_total_murder <- murder$population[i_min]

kable(rbind(max_total_murder,min_total_murder,
      State_max_total_murdre, state_min_total_murdre,
      Population_max_total_murder, Population_min_total_murder),
      "pandoc",
      caption = "The first six rows of murder dataset with adding       murder rate column",
      align = "c",
      font_size = 5)
```

Table 6: The first six rows of murder dataset with adding murder rate column

max_total_murder	1257
min_total_murder	2
State_max_total_murdre	California
state_min_total_murdre	Vermont
Population_max_total_murder	37253956
Population_min_total_murder	625741

In order to compare the total murder of each state, it is better defined one variable as murder rate. Murder rate is defined as the total murder to 100000 population of each state. The first six rows of murder dataset is shown in Table 7.

```r
murder <-murder %>% mutate(region = factor(region),
                           murder_rate = total / population *                    10^5)
kable(head(murder),
      "pandoc",
      caption = "The first six rows of murder dataset with adding murder rate column",
      align = "c",
      font_size = 5)
```

Table 7: The first six rows of murder dataset with adding murder rate column

state	abb	region	population	total	murder_rate
Alabama	AL	South	4779736	135	2.824424
Alaska	AK	West	710231	19	2.675186

state	abb	region	population	total	murder_rate
Arizona	AZ	West	6392017	232	3.629527
Arkansas	AR	South	2915918	93	3.189390
California	CA	West	37253956	1257	3.374138
Colorado	CO	West	5029196	65	1.292453

Based on murder rate column, the maximum murder rate is 16.42 per 100000 population for District of Columbia.Also the minimum murder rate is 0.32 per 100000 population for Vermont (Table 8).

```
# The state name of min and max of murder rate
max_murder_rate <- max(murder$murder_rate)
min_murder_rate <- min(murder$murder_rate)

i_max_murder_rate <- which.max(murder$murder_rate)
State_max_murdre_rate <- murder$state[i_max_murder_rate]
i_min_murder_rate  <- which.min(murder$murder_rate)
state_min_murdre_rate <- murder$state[i_min_murder_rate ]

kable(rbind(max_murder_rate,min_murder_rate,
      State_max_murdre_rate, state_min_murdre_rate),
      "pandoc",
      caption = "The first six rows of murder dataset with adding murder rate column",
      align = "c",
      font_size = 5)
```

Table 8: The first six rows of murder dataset with adding murder rate column

max_murder_rate	16.4527531771264
min_murder_rate	0.319621057274495
State_max_murdre_rate	District of Columbia
state_min_murdre_rate	Vermont

The distribution of murder rate per 100000 population is shown for each state in Figure 1.

```
murder %>% mutate(abb = reorder(abb,murder_rate)) %>%
  ggplot(aes(abb, murder_rate)) +
  geom_col(width = 0.7, color = "pink",fill = "blue") +
  xlab("Abb") +
  ylab("Murder rate per 100000" ) +
  coord_flip() +
  theme(axis.text.x = element_text(size = 10)) +
  theme(axis.text.y = element_text(size = 4))
```

The histogram of murder rate shows that there is a wide range of values with most of them between 2 and 3 and one very extreme case with a murder rate of more than 15 (Figure 2).

The murder rate and population are compared in different region as well. Figure 3 shows that in South region the murder rate and total murder are higher than other region. although these relation is not correct for other regions.

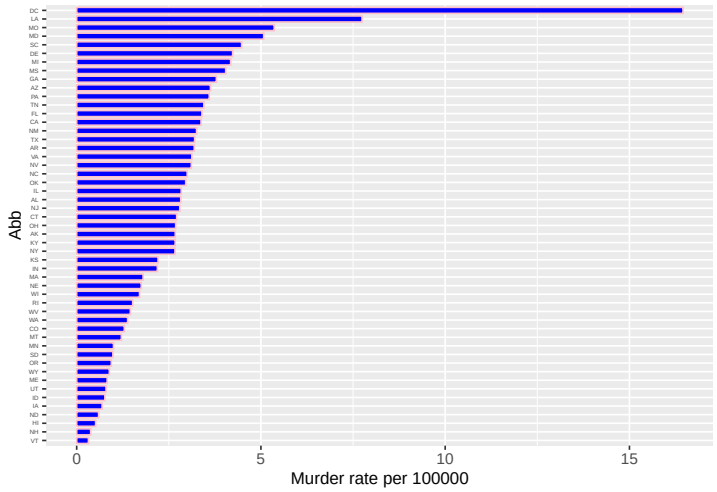

Figure 1: Variation of murder rate based on state

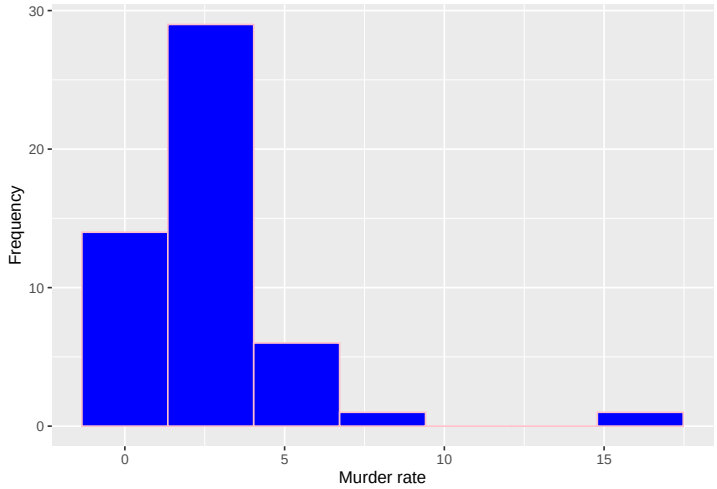

Figure 2: Distribution of murder rate

```
p1_murder_rate <- murder %>% group_by(region) %>%
  ggplot(aes(region, murder_rate)) +
  geom_boxplot(col = "pink", fill = "blue")+
  xlab("Region") + ylab("Murder rate")

p1_population <- murder %>%
  group_by(region) %>%
  ggplot(aes(region, population/10^6)) +
  geom_bar(aes(fill = region ),stat="identity")+
  xlab("Region") +
  ylab("Population in million")+
  theme(legend.position =  "none") +
  scale_y_log10()

grid.arrange(p1_murder_rate,p1_population)
```

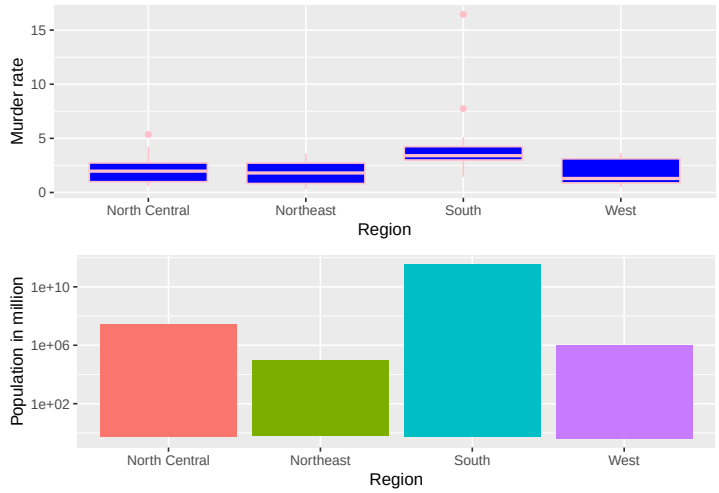

Figure 3: Varioation of murder rate and total murder versus region

The total murder and murder rate of each state versus population in million at log scale are shown detailed with state at different region (Figure 5).

```
p1 <- murder %>% ggplot(aes(population/10^6, total, label = abb)) +
  geom_point(aes(col=region), size = 3) +
  geom_text_repel() +
  scale_x_log10() +
  scale_y_log10() +
  xlab("Populations in millions (log scale)") +
  ylab("Total murders (log scale)") +
  ggtitle("Total Murder versus population in millions") +
  scale_color_discrete(name = "Region")+
  facet_grid(. ~ region) +
  theme(legend.position = "none")

p2 <- murder %>% ggplot(aes(population/10^6, murder_rate, label = abb)) +
  geom_point(aes(col=region), size = 3) +
  geom_text_repel() +
  scale_x_log10() +
```

```
  scale_y_log10() +
  xlab("Populations in millions (log scale)") +
  ylab("Murder rate (log scale)") +
  ggtitle("Murder rate versus Population in millions") +
  scale_color_discrete(name = "Region")+
  facet_grid(. ~ region) +
  theme(legend.position = "none")

grid.arrange(p1,p2)
```

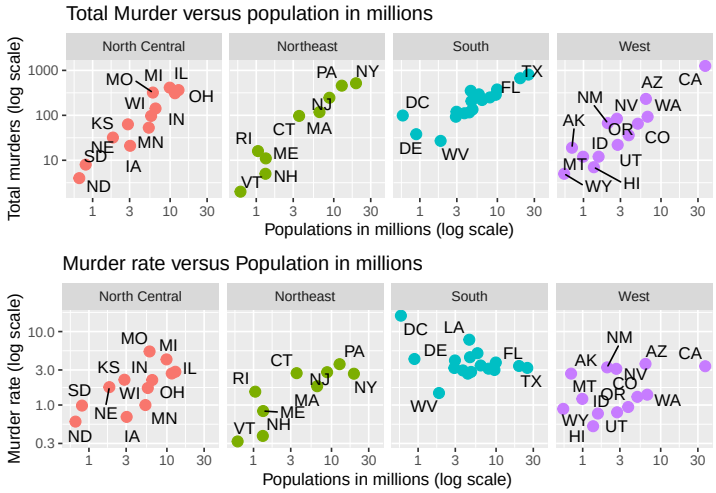

Figure 4: Varioation of Murder rate (log scale) and Total murder (Log scale) versus population in million (log scale)