Final Project Proposal

Hailie Mathews

2024-03-09

```r
library(openintro)
```

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata
```

```r
library(dplyr)
```

```
##
## 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
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(usdata)
library(ggplot2)
library(maps)
```

```
##
## Attaching package: 'maps'
##
## The following object is masked from 'package:purrr':
##
##     map
```

“Introduction”

Focus will be on the **state_stats** dataset. The main question I will be exploring is, ***“What connection, if any, exists between population size (2010 only), poverty, and robbery?”***

My hypothesis: *There is a correlation between high population, high poverty levels and high robbery rates.*

I got the dataset from OpenIntro datasets, under **usdata**. **State_stats** is one of 13 datasets inside of **usdata**.

Some information about OpenIntro datasets: *Directly from their site, it says, “Supplemental functions and data for OpenIntro resources, which includes open-source textbooks and resources for introductory statistics.”*

The data was collected from the official US Census website along with various other sources.

There are 24 variables and 51 observations in the dataset. Each variable represents a different category, relating to many things, ranging from population to income to federal spending. Each observation represents the state that the data was collected from.

Along with population, poverty, and robbery, I will also be looking at individuals collecting social security, federal spending, and income per capita. These will not effect my conclusions regarding the dataset but I found them interesting to explore alongside my main question.

“Data”

Data Dimensions:

```r
colnames(state_stats)
```

```
##  [1] "state"            "abbr"             "fips"             "pop2010"
##  [5] "pop2000"          "homeownership"    "multiunit"        "income"
##  [9] "med_income"       "poverty"          "fed_spend"        "land_area"
## [13] "smoke"            "murder"           "robbery"          "agg_assault"
## [17] "larceny"          "motor_theft"      "soc_sec"          "nuclear"
## [21] "coal"             "tr_deaths"        "tr_deaths_no_alc" "unempl"
```

Codebook:

state: State name.

abbr: State abbreviation (e.g. “MN”).

fips: FIPS code.

pop2010: Population in 2010.

pop2000: Population in 2000.

homeownership: Home ownership rate.

multiunit: Percent of living units that are in multi-unit structures.

income: Average income per capita.

med_income: Median household income.

poverty: Poverty rate.

fed_spend: Federal spending per capita.

land_area: Land area.

smoke: Percent of population that smokes.

murder: Murders per 100,000 people.

robbery: Robberies per 100,000.

agg_assault: Aggravated assaults per 100,000.

larceny: Larcenies per 100,000.

motor_theft: Vehicle theft per 100,000.

soc_sec: Percent of individuals collecting social security.

nuclear: Percent of power coming from nuclear sources.

coal: Percent of power coming from coal sources.

tr_deaths: Traffic deaths per 100,000.

tr_deaths_no_alc: Traffic deaths per 100,000 where alcohol was not a factor.

unempl: Unemployment rate (February 2012, preliminary).

“Data Analysis Plan”

Outcome variables will be *poverty* and *robbery*.

Predictor variable will be *pouplation size*.

No comparison groups, as of now.

Robbery and Population datasets:

```r
state_stats %>%
  count(abbr, robbery, pop2010) %>%
  group_by(pop2010) %>%
  filter(robbery >= 0) %>%
  arrange(desc(robbery))
```

```
## # A tibble: 48 x 4
## # Groups:   pop2010 [48]
##    abbr  robbery  pop2010     n
##    <fct>   <dbl>    <dbl> <int>
##  1 DC       672.   601723     1
##  2 MD       257.  5773552     1
##  3 NV       195.  2700551     1
##  4 NY       183. 19378102     1
##  5 IL       182. 12830632     1
##  6 CA       176. 37253956     1
##  7 FL       169. 18801310     1
##  8 TN       167.  6346105     1
##  9 OH       163. 11536504     1
## 10 TX       157. 25145561     1
## # i 38 more rows
```

```r
state_stats %>%
  count(abbr, robbery, pop2010) %>%
  group_by(pop2010) %>%
  filter(robbery >= 0) %>%
  arrange(desc(pop2010))
```

```
## # A tibble: 48 x 4
## # Groups:   pop2010 [48]
##    abbr  robbery  pop2010     n
##    <fct>   <dbl>    <dbl> <int>
##   1 CA        176. 37253956     1
##   2 TX        157. 25145561     1
##   3 NY        183. 19378102     1
##   4 FL        169. 18801310     1
##   5 IL        182. 12830632     1
##   6 PA        155. 12702379     1
##   7 OH        163. 11536504     1
##   8 MI        132.  9883640     1
##   9 GA        155.  9687653     1
##  10 NC        146.  9535483     1
## # i 38 more rows
```

Poverty and Population datasets:

```
state_stats %>%
  count(abbr, poverty, pop2010) %>%
  group_by(pop2010) %>%
  filter(poverty >= 0) %>%
  arrange(desc(poverty))
```

```
## # A tibble: 51 x 4
## # Groups:   pop2010 [51]
##    abbr  poverty  pop2010     n
##    <fct>    <dbl>    <dbl> <int>
##   1 MS        21.2  2967297     1
##   2 DC        18.5   601723     1
##   3 NM        18.4  2059179     1
##   4 LA        18.1  4533372     1
##   5 AR        18    2915918     1
##   6 KY        17.7  4339367     1
##   7 WV        17.4  1852994     1
##   8 AL        17.1  4779736     1
##   9 TX        16.8 25145561     1
##  10 TN        16.5  6346105     1
## # i 41 more rows
```

```
state_stats %>%
  count(abbr, poverty, pop2010) %>%
  group_by(pop2010) %>%
  filter(poverty >= 0) %>%
  arrange(desc(pop2010))
```

```
## # A tibble: 51 x 4
## # Groups:   pop2010 [51]
##    abbr  poverty  pop2010     n
##    <fct>    <dbl>    <dbl> <int>
##   1 CA        13.7 37253956     1
##   2 TX        16.8 25145561     1
##   3 NY        14.2 19378102     1
##   4 FL        13.8 18801310     1
##   5 IL        12.6 12830632     1
##   6 PA        12.4 12702379     1
##   7 OH        14.2 11536504     1
##   8 MI        14.8  9883640     1
##   9 GA        15.7  9687653     1
##  10 NC        15.5  9535483     1
## # i 41 more rows
```

Methods include: separating data from high to low, comparing the three variables (as seen above) and observing any correlation.

A substantial observed correlation between population, poverty and robbery will be needed to support my hypothesized answer.