

# P8130 Final Project

## Abstract

## Introduction (brief context and background of the problem)

## Methods (data description and statistical methods)

## Results

## Conclusions/Discussion

```
library(tidyverse)
library(PerformanceAnalytics)
```

## Read in dataset

```
cdi = read_csv("./cdi.csv") %>%
  janitor::clean_names()

## no missing value
cdi %>%
  select(everything()) %>%
  summarise_all(funs(sum(is.na(.)))) %>%
  knitr::kable()
```

id	cty	state	area	pop	pop18	pop65	docs	beds	crimes	hsgrad	bagrad	poverty	unempp	pcincome	totalinc	region
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

## Data cleaning

```
# some normalization for better comparison
cdi =
  cdi %>%
  mutate(crm_1000 = crimes/pop*1000, # as indicated by the project prompt
         docs_1000 = docs/pop*1000, # every 1000 people how many doctors
         beds_1000 = beds/pop*1000, # ratio of beds per doctor, indicating a
         pop_density = pop/area, # how many people per square miles
         region = factor(region)) %>%
  select(-id, -cty, -crimes, -area, -docs, -beds)
```

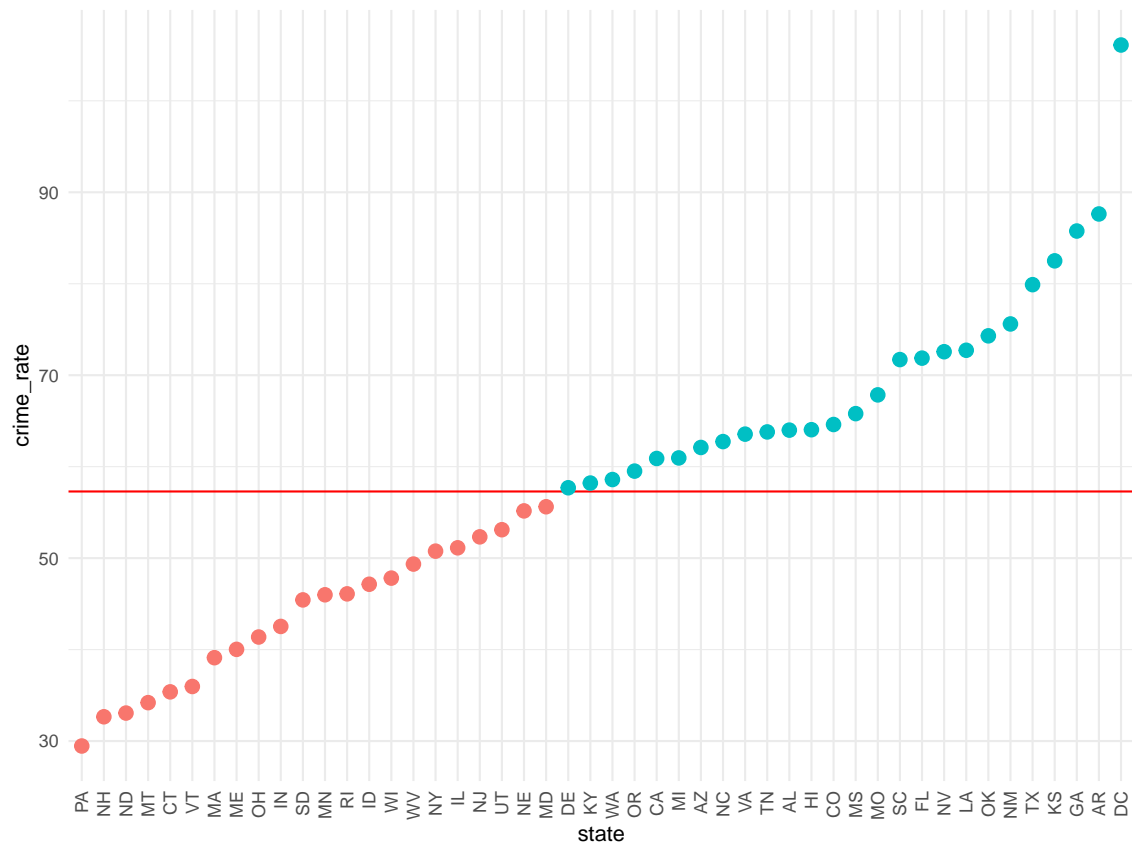
## Data Exploration

```
## summary statistics, tentative, NOT FINAL
sum_cdi =
  cdi %>%
  select(crm_1000, docs_1000, pop_density, pop, pop18, pop65, hsgrad, bagrad, poverty, unemp, pcincome, totalinc, region)
summary(sum_cdi)
```

```
##      crm_1000      docs_1000      pop_density      pop
## Min.   : 4.601   Min.   : 0.3559   Min.   : 13.26   Min.   : 100043
## 1st Qu.: 38.102  1st Qu.: 1.2127   1st Qu.: 192.34  1st Qu.: 139027
## Median : 52.429  Median : 1.7509   Median : 335.91  Median : 217280
## Mean   : 57.286  Mean   : 2.1230   Mean   : 888.44  Mean   : 393011
## 3rd Qu.: 72.597  3rd Qu.: 2.4915   3rd Qu.: 756.55  3rd Qu.: 436064
## Max.   :295.987  Max.   :17.0377   Max.   :32403.72  Max.   :8863164
##      pop18      pop65      hsgrad      bagrad
## Min.   :16.40   Min.   : 3.000   Min.   :46.60   Min.   : 8.10
## 1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:73.88   1st Qu.:15.28
## Median :28.10   Median :11.750   Median :77.70   Median :19.70
## Mean   :28.57   Mean   :12.170   Mean   :77.56   Mean   :21.08
## 3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.:82.40   3rd Qu.:25.32
## Max.   :49.70   Max.   :33.800   Max.   :92.90   Max.   :52.30
##      poverty      unemp      pcincome      beds_1000
## Min.   : 1.400   Min.   : 2.200   Min.   : 8899   Min.   : 0.1649
## 1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118   1st Qu.: 2.1972
## Median : 7.900   Median : 6.200   Median :17759   Median : 3.3287
## Mean   : 8.721   Mean   : 6.597   Mean   :18561   Mean   : 3.6493
## 3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270   3rd Qu.: 4.5649
## Max.   :36.300   Max.   :21.300   Max.   :37541   Max.   :19.6982
```

```
mean_crm = mean(cdi$crm_1000)
cdi_state = cdi %>%
  group_by(state) %>%
  summarize(crime_rate = mean(crm_1000)) %>%
  mutate(low_high = ifelse(crime_rate>mean_crm, TRUE,FALSE))

cdi_state %>%
  mutate(state = fct_reorder(state, crime_rate)) %>%
  ggplot(aes(x = state, y = crime_rate))+
  geom_hline(yintercept = mean_crm, color = "red")+
  geom_point(aes(color = low_high),size = 3)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust= 1),
        legend.position = "none")
```



```
corr_matrix =
  cdi %>%
  select(-state, -region) %>%
  chart.Correlation(histogram = TRUE, method = "pearson")
```

