P8130 Final Project

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	unemp	pcincome	totalinc	region
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

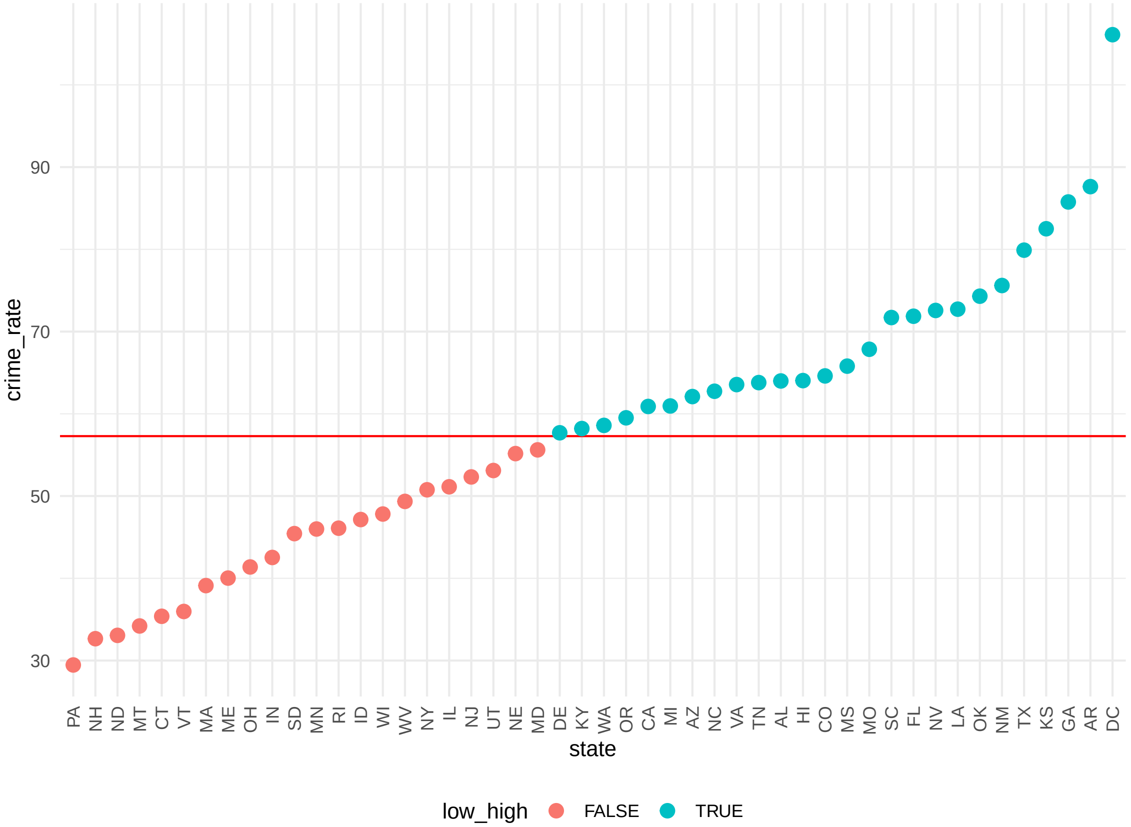
Data cleaning

```
cdi =
  cdi %>%
  mutate(crm_1000 = crimes/pop*1000,  # as indicated by the project prompt
         docs_rate_1000 = docs/pop*1000,  # every 1000 people how many doctors
         beds_docs = beds/docs,
         region = factor(region)) %>%
  select(-id, -cty, -crimes)
cdi
```

```
## # A tibble: 440 x 17
##    state  area     pop pop18 pop65  docs  beds hsgrad bagrad poverty unemp
##    <chr> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>   <dbl> <dbl>
##  1 CA     4060 8863164  32.1   9.7 23677 27700   70     22.3    11.6   8
##  2 IL      946 5105067  29.2  12.4 15153 21550   73.4   22.8    11.1   7.2
##  3 TX     1729 2818199  31.3   7.1  7553 12449   74.9   25.4    12.5   5.7
##  4 CA     4205 2498016  33.5  10.9  5905  6179   81.9   25.3     8.1   6.1
##  5 CA      790 2410556  32.6   9.2  6062  6369   81.2   27.8     5.2   4.8
##  6 NY       71 2300664  28.3  12.4  4861  8942   63.7   16.6    19.5   9.5
##  7 AZ     9204 2122101  29.2  12.5  4320  6104   81.5   22.1     8.8   4.9
##  8 MI      614 2111687  27.4  12.5  3823  9490   70     13.7    16.9  10
##  9 FL     1945 1937094  27.1  13.9  6274  8840   65     18.8    14.2   8.7
## 10 TX      880 1852810  32.6   8.2  4718  6934   77.1   26.3    10.4   6.1
## # ... with 430 more rows, and 6 more variables: pcincome <dbl>, totalinc <dbl>,
## #   region <fct>, crm_1000 <dbl>, docs_rate_1000 <dbl>, beds_docs <dbl>
```

```
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))
```

```r
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))
```



Data Exploration

```r
## summary statistics
knitr::kable(summary(cdi))
```

	state	area	pop	pop18	pop65	docs	beds	hsgrad	bagrad	poverty	unemp	pcincome	totalinc	region	crm_1000	docs_rate_1000	beds_rate_1000
	Length:440	Min. : 15.0	Min. : 100043	Min. :16.40	Min. : 3.000	Min. : 39.0	Min. : 92.0	Min. :46.60	Min. : 8.10	Min. : 1.400	Min. : 2.200	Min. : 8899	Min. : 1141	1:103	Min. : 4.601	Min. : 0.3559	Min. :0.07969
	Class :character	1st Qu.: 451.2	1st Qu.: 139027	1st Qu.:26.20	1st Qu.: 9.875	1st Qu.: 182.8	1st Qu.: 390.8	1st Qu.:73.88	1st Qu.:15.28	1st Qu.: 5.300	1st Qu.: 5.100	1st Qu.:16118	1st Qu.: 2311	2:108	1st Qu.: 38.102	1st Qu.: 1.2127	1st Qu.:1.34565
	Mode :character	Median : 656.5	Median : 217280	Median :28.10	Median :11.750	Median : 401.0	Median : 755.0	Median :77.70	Median :19.70	Median : 7.900	Median : 6.200	Median :17759	Median : 3857	3:152	Median : 52.429	Median : 1.7509	Median :1.83419

state	area	pop	pop18	pop65	docs	beds	hsgrad	bagrad	poverty	unemp	pcincome	totalinc	region	crm_1000	docs_rate_1000	beds_docs
NA	Mean : 1041.4	Mean : 393011	Mean :28.57	Mean :12.170	Mean : 988.0	Mean : 1458.6	Mean :77.56	Mean :21.08	Mean : 8.721	Mean : 6.597	Mean :18561	Mean : 7869	4: 77	Mean : 57.286	Mean : 2.1230	Mean :1.97855
NA	3rd Qu.: 946.8	3rd Qu.: 436064	3rd Qu.:30.02	3rd Qu.:13.625	3rd Qu.: 1036.0	3rd Qu.: 1575.8	3rd Qu.:82.40	3rd Qu.:25.32	3rd Qu.:10.900	3rd Qu.: 7.500	3rd Qu.:20270	3rd Qu.: 8654	NA	3rd Qu.: 72.597	3rd Qu.: 2.4915	3rd Qu.:2.42710
NA	Max. :20062.0	Max. :8863164	Max. :49.70	Max. :33.800	Max. :23677.0	Max. :27700.0	Max. :92.90	Max. :52.30	Max. :36.300	Max. :21.300	Max. :37541	Max. :184230	NA	Max. :295.987	Max. :17.0377	Max. :7.41667

```
cdi
```

```
## # A tibble: 440 x 17
##    state  area     pop pop18 pop65  docs  beds hsgrad bagrad poverty unemp
##    <chr> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>   <dbl> <dbl>
##  1 CA     4060 8863164  32.1   9.7 23677 27700   70     22.3    11.6   8
##  2 IL      946 5105067  29.2  12.4 15153 21550   73.4   22.8    11.1   7.2
##  3 TX     1729 2818199  31.3   7.1  7553 12449   74.9   25.4    12.5   5.7
##  4 CA     4205 2498016  33.5  10.9  5905  6179   81.9   25.3     8.1   6.1
##  5 CA      790 2410556  32.6   9.2  6062  6369   81.2   27.8     5.2   4.8
##  6 NY       71 2300664  28.3  12.4  4861  8942   63.7   16.6    19.5   9.5
##  7 AZ     9204 2122101  29.2  12.5  4320  6104   81.5   22.1     8.8   4.9
##  8 MI      614 2111687  27.4  12.5  3823  9490   70     13.7    16.9  10
##  9 FL     1945 1937094  27.1  13.9  6274  8840   65     18.8    14.2   8.7
## 10 TX      880 1852810  32.6   8.2  4718  6934   77.1   26.3    10.4   6.1
## # ... with 430 more rows, and 6 more variables: pcincome <dbl>, totalinc <dbl>,
## #   region <fct>, crm_1000 <dbl>, docs_rate_1000 <dbl>, beds_docs <dbl>
```