P8130 Final Project

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(pastecs)
```

```
##
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':
##
##     first, last

## The following object is masked from 'package:tidyr':
##
##     extract
```

Read in dataset

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

```
## Rows: 440 Columns: 17

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): cty, state
## dbl (15): id, area, pop, pop18, pop65, docs, beds, crimes, hsgrad, bagrad, p...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
cdi
```

```
## # A tibble: 440 x 17
##       id cty     state  area    pop pop18 pop65  docs  beds crimes hsgrad bagrad
##    <dbl> <chr>   <chr> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>
##  1     1 Los_An~ CA     4060 8.86e6  32.1   9.7 23677 27700 688936   70     22.3
##  2     2 Cook    IL      946 5.11e6  29.2  12.4 15153 21550 436936   73.4   22.8
##  3     3 Harris  TX     1729 2.82e6  31.3   7.1  7553 12449 253526   74.9   25.4
##  4     4 San_Di~ CA     4205 2.50e6  33.5  10.9  5905  6179 173821   81.9   25.3
##  5     5 Orange  CA      790 2.41e6  32.6   9.2  6062  6369 144524   81.2   27.8
##  6     6 Kings   NY       71 2.30e6  28.3  12.4  4861  8942 680966   63.7   16.6
##  7     7 Marico~ AZ     9204 2.12e6  29.2  12.5  4320  6104 177593   81.5   22.1
##  8     8 Wayne   MI      614 2.11e6  27.4  12.5  3823  9490 193978   70     13.7
##  9     9 Dade    FL     1945 1.94e6  27.1  13.9  6274  8840 244725   65     18.8
## 10    10 Dallas  TX      880 1.85e6  32.6   8.2  4718  6934 214258   77.1   26.3
## # ... with 430 more rows, and 5 more variables: poverty <dbl>, unemp <dbl>,
## #   pcincome <dbl>, totalinc <dbl>, region <dbl>
```

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

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
## # A tibble: 1 x 17
##      id   cty state  area   pop pop18 pop65  docs  beds crimes hsgrad bagrad
##   <int> <int> <int> <int> <int> <int> <int> <int> <int>  <int>  <int>  <int>
## 1     0     0     0     0     0     0     0     0     0      0      0      0
## # ... with 5 more variables: poverty <int>, unemp <int>, pcincome <int>,
## #   totalinc <int>, region <int>
```

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)
cdi
```

```
## # A tibble: 440 x 20
##       id cty     state  area    pop pop18 pop65  docs  beds crimes hsgrad bagrad
##    <dbl> <chr>   <chr> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>
##  1     1 Los_An~ CA     4060 8.86e6  32.1   9.7 23677 27700 688936   70     22.3
##  2     2 Cook    IL      946 5.11e6  29.2  12.4 15153 21550 436936   73.4   22.8
##  3     3 Harris  TX     1729 2.82e6  31.3   7.1  7553 12449 253526   74.9   25.4
##  4     4 San_Di~ CA     4205 2.50e6  33.5  10.9  5905  6179 173821   81.9   25.3
##  5     5 Orange  CA      790 2.41e6  32.6   9.2  6062  6369 144524   81.2   27.8
##  6     6 Kings   NY       71 2.30e6  28.3  12.4  4861  8942 680966   63.7   16.6
##  7     7 Marico~ AZ     9204 2.12e6  29.2  12.5  4320  6104 177593   81.5   22.1
##  8     8 Wayne   MI      614 2.11e6  27.4  12.5  3823  9490 193978   70     13.7
## 9      9 Dade    FL      1945 1.94e6  27.1  13.9  6274  8840 244725   65     18.8
## 10    10 Dallas  TX       880 1.85e6  32.6   8.2  4718  6934 214258   77.1   26.3
## # ... with 430 more rows, and 8 more variables: poverty <dbl>, unemp <dbl>,
## #   pcincome <dbl>, totalinc <dbl>, region <dbl>, crm_1000 <dbl>,
## #   docs_rate_1000 <dbl>, beds_docs <dbl>
```

Data Exploration

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

id	cty	state	area	pop	pop18	pop65	docs	beds	crimes	hsgrad	bagrad	poverty	unemp	pcincome	totalinc	region	crm_1000	docs_rate_1000	beds_docs
Min. : 1.0	Length:440	Length:440	Min. : 15.0	Min. : 100043	Min. :16.40	Min. : 3.000	Min. : 39.0	Min. : 92.0	Min. : 563	Min. :46.60	Min. : 8.10	Min. : 1.400	Min. : 2.200	Min. : 8899	Min. : 1141	Min. :1.000	Min. : 4.601	Min. : 0.3559	Min. :0.07969
1st Qu.:110.8	Class :character	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.: 6220	1st Qu.:73.88	1st Qu.:15.28	1st Qu.: 5.300	1st Qu.: 5.100	1st Qu.:16118	1st Qu.: 2311	1st Qu.:2.000	1st Qu.: 38.102	1st Qu.: 1.2127	1st Qu.:1.34565
Median :220.5	Mode :character	Mode :character	Median : 656.5	Median : 217280	Median :28.10	Median :11.750	Median : 401.0	Median : 755.0	Median : 11820	Median :77.70	Median :19.70	Median : 7.900	Median : 6.200	Median :17759	Median : 3857	Median :3.000	Median : 52.429	Median : 1.7509	Median :1.83419
Mean :220.5	NA	NA	Mean : 1041.4	Mean : 393011	Mean :28.57	Mean :12.170	Mean : 988.0	Mean : 1458.6	Mean : 27112	Mean :77.56	Mean :21.08	Mean : 8.721	Mean : 6.597	Mean :18561	Mean : 7869	Mean :2.461	Mean : 57.286	Mean : 2.1230	Mean :1.97855
3rd Qu.:330.2	NA	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.: 26280	3rd Qu.:82.40	3rd Qu.:25.32	3rd Qu.:10.900	3rd Qu.: 7.500	3rd Qu.:20270	3rd Qu.: 8654	3rd Qu.:3.000	3rd Qu.: 72.597	3rd Qu.: 2.4915	3rd Qu.:2.42710
Max. :440.0	NA	NA	Max. :20062	Max. :8863164	Max. :49.70	Max. :33.80	Max. :23677	Max. :27700	Max. :688936	Max. :92.90	Max. :52.30	Max. :36.30	Max. :21.30	Max. :37541	Max. :184230	Max. :4.000	Max. :295.987	Max. :17.0377	Max. :7.741667

```
cdi
```

```
## # A tibble: 440 x 20
##       id cty     state  area    pop pop18 pop65  docs  beds crimes hsgrad bagrad
##    <dbl> <chr>   <chr> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>
## 1      1 Los_An~ CA     4060 8.86e6  32.1   9.7 23677 27700 688936   70     22.3
## 2      2 Cook    IL      946 5.11e6  29.2  12.4 15153 21550 436936   73.4   22.8
## 3      3 Harris  TX     1729 2.82e6  31.3   7.1  7553 12449 253526   74.9   25.4
## 4      4 San_Di~ CA     4205 2.50e6  33.5  10.9  5905  6179 173821   81.9   25.3
## 5      5 Orange  CA      790 2.41e6  32.6   9.2  6062  6369 144524   81.2   27.8
## 6      6 Kings   NY       71 2.30e6  28.3  12.4  4861  8942 680966   63.7   16.6
## 7      7 Marico~ AZ     9204 2.12e6  29.2  12.5  4320  6104 177593   81.5   22.1
## 8      8 Wayne   MI      614 2.11e6  27.4  12.5  3823  9490 193978   70     13.7
## 9      9 Dade    FL     1945 1.94e6  27.1  13.9  6274  8840 244725   65     18.8
## 10    10 Dallas  TX      880 1.85e6  32.6   8.2  4718  6934 214258   77.1   26.3
## # ... with 430 more rows, and 8 more variables: poverty <dbl>, unemp <dbl>,
## #   pcincome <dbl>, totalinc <dbl>, region <dbl>, crm_1000 <dbl>,
## #   docs_rate_1000 <dbl>, beds_docs <dbl>
```