Lab_01-1

Emmenta Janneh

2024-01-17

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- 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 errors
```

Lab 01-1

Load the CDI dataset into R, saving the data frame in a variable called `cdi`.

```
cdi <- read_csv("https://dcgerard.github.io/stat_415_615/data/cdi.csv")
```

```
## Rows: 440 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): county, state, region
## dbl (14): id, area, pop, percent_18_34, percent_65, physicians, beds, crimes...
##
## 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.
```

Use a function that provides a rough glimpse at these data.

```
glimpse(cdi)
```

```
## Rows: 440
## Columns: 17
## $ id            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ county        <chr> "Los_Angeles", "Cook", "Harris", "San_Diego", "Orange", ~
```

```
## $ state         <chr> "CA", "IL", "TX", "CA", "CA", "NY", "AZ", "MI", "FL", "T~
## $ area          <dbl> 4060, 946, 1729, 4205, 790, 71, 9204, 614, 1945, 880, 13~
## $ pop           <dbl> 8863164, 5105067, 2818199, 2498016, 2410556, 2300664, 21~
## $ percent_18_34 <dbl> 32.1, 29.2, 31.3, 33.5, 32.6, 28.3, 29.2, 27.4, 27.1, 32~
## $ percent_65    <dbl> 9.7, 12.4, 7.1, 10.9, 9.2, 12.4, 12.5, 12.5, 13.9, 8.2, ~
## $ physicians    <dbl> 23677, 15153, 7553, 5905, 6062, 4861, 4320, 3823, 6274, ~
## $ beds          <dbl> 27700, 21550, 12449, 6179, 6369, 8942, 6104, 9490, 8840,~
## $ crimes        <dbl> 688936, 436936, 253526, 173821, 144524, 680966, 177593, ~
## $ high_school   <dbl> 70.0, 73.4, 74.9, 81.9, 81.2, 63.7, 81.5, 70.0, 65.0, 77~
## $ bachelors     <dbl> 22.3, 22.8, 25.4, 25.3, 27.8, 16.6, 22.1, 13.7, 18.8, 26~
## $ poverty       <dbl> 11.6, 11.1, 12.5, 8.1, 5.2, 19.5, 8.8, 16.9, 14.2, 10.4,~
## $ unemployment  <dbl> 8.0, 7.2, 5.7, 6.1, 4.8, 9.5, 4.9, 10.0, 8.7, 6.1, 8.0, ~
## $ capita_income <dbl> 20786, 21729, 19517, 19588, 24400, 16803, 18042, 17461, ~
## $ total_income  <dbl> 184230, 110928, 55003, 48931, 58818, 38658, 38287, 36872~
## $ region        <chr> "W", "NC", "S", "W", "W", "NE", "W", "NC", "S", "S", "NE~
```

Calculate the mean and standard deviation of the population of the counties

```
pop_mean <- mean(cdi$pop)
pop_mean
```

```
## [1] 393010.9
```

```
pop_sd <- sd(cdi$pop)
pop_sd
```

```
## [1] 601987
```

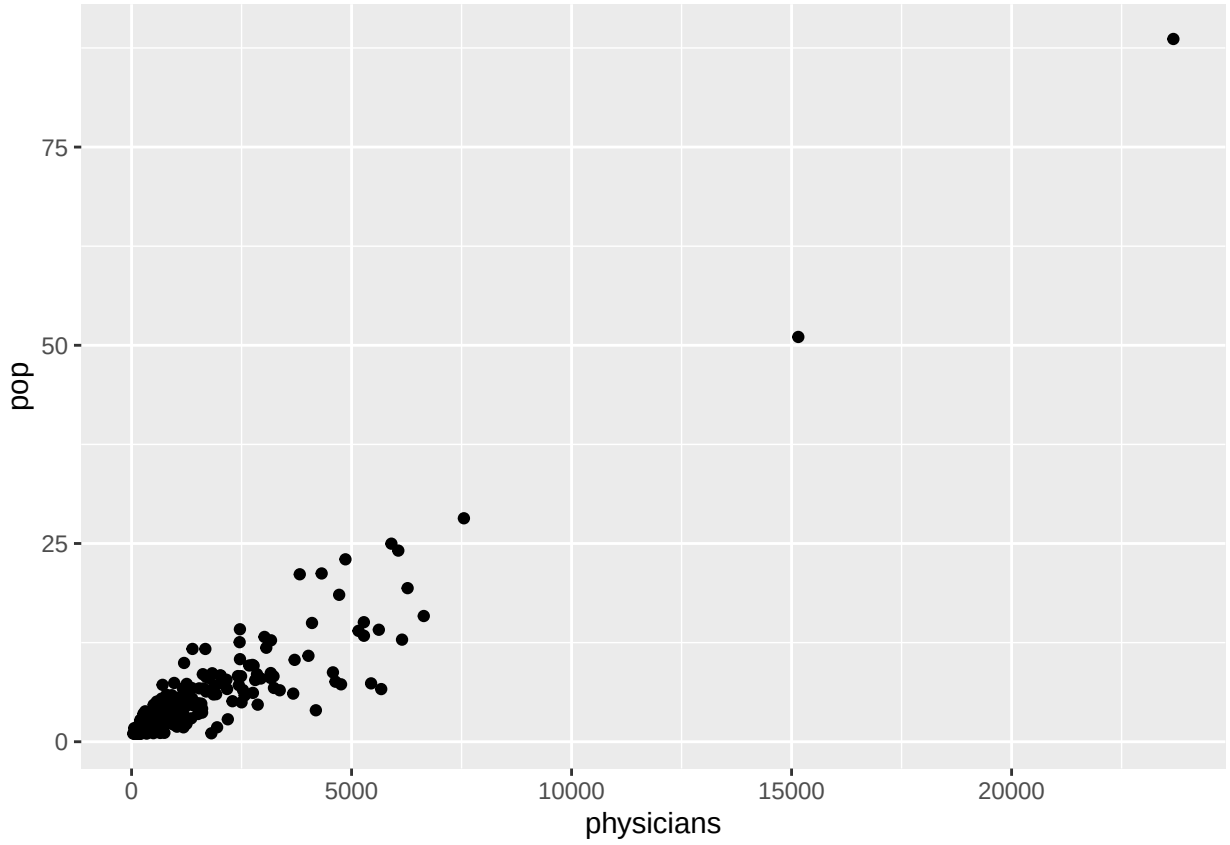
Convert the pop variable to be in units of 100,000 individuals (so 1 corresponds to 100,000, 2 corresponds to 200,000, ect).

```
cdi <- mutate(cdi, pop = pop/100000)
glimpse(cdi)
```

```
## Rows: 440
## Columns: 17
## $ id            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ county        <chr> "Los_Angeles", "Cook", "Harris", "San_Diego", "Orange", ~
## $ state         <chr> "CA", "IL", "TX", "CA", "CA", "NY", "AZ", "MI", "FL", "T~
## $ area          <dbl> 4060, 946, 1729, 4205, 790, 71, 9204, 614, 1945, 880, 13~
## $ pop           <dbl> 88.63164, 51.05067, 28.18199, 24.98016, 24.10556, 23.006~
## $ percent_18_34 <dbl> 32.1, 29.2, 31.3, 33.5, 32.6, 28.3, 29.2, 27.4, 27.1, 32~
## $ percent_65    <dbl> 9.7, 12.4, 7.1, 10.9, 9.2, 12.4, 12.5, 12.5, 13.9, 8.2, ~
## $ physicians    <dbl> 23677, 15153, 7553, 5905, 6062, 4861, 4320, 3823, 6274, ~
## $ beds          <dbl> 27700, 21550, 12449, 6179, 6369, 8942, 6104, 9490, 8840,~
## $ crimes        <dbl> 688936, 436936, 253526, 173821, 144524, 680966, 177593, ~
## $ high_school   <dbl> 70.0, 73.4, 74.9, 81.9, 81.2, 63.7, 81.5, 70.0, 65.0, 77~
## $ bachelors     <dbl> 22.3, 22.8, 25.4, 25.3, 27.8, 16.6, 22.1, 13.7, 18.8, 26~
## $ poverty       <dbl> 11.6, 11.1, 12.5, 8.1, 5.2, 19.5, 8.8, 16.9, 14.2, 10.4,~
## $ unemployment  <dbl> 8.0, 7.2, 5.7, 6.1, 4.8, 9.5, 4.9, 10.0, 8.7, 6.1, 8.0, ~
## $ capita_income <dbl> 20786, 21729, 19517, 19588, 24400, 16803, 18042, 17461, ~
## $ total_income  <dbl> 184230, 110928, 55003, 48931, 58818, 38658, 38287, 36872~
## $ region        <chr> "W", "NC", "S", "W", "W", "NE", "W", "NC", "S", "S", "NE~
```

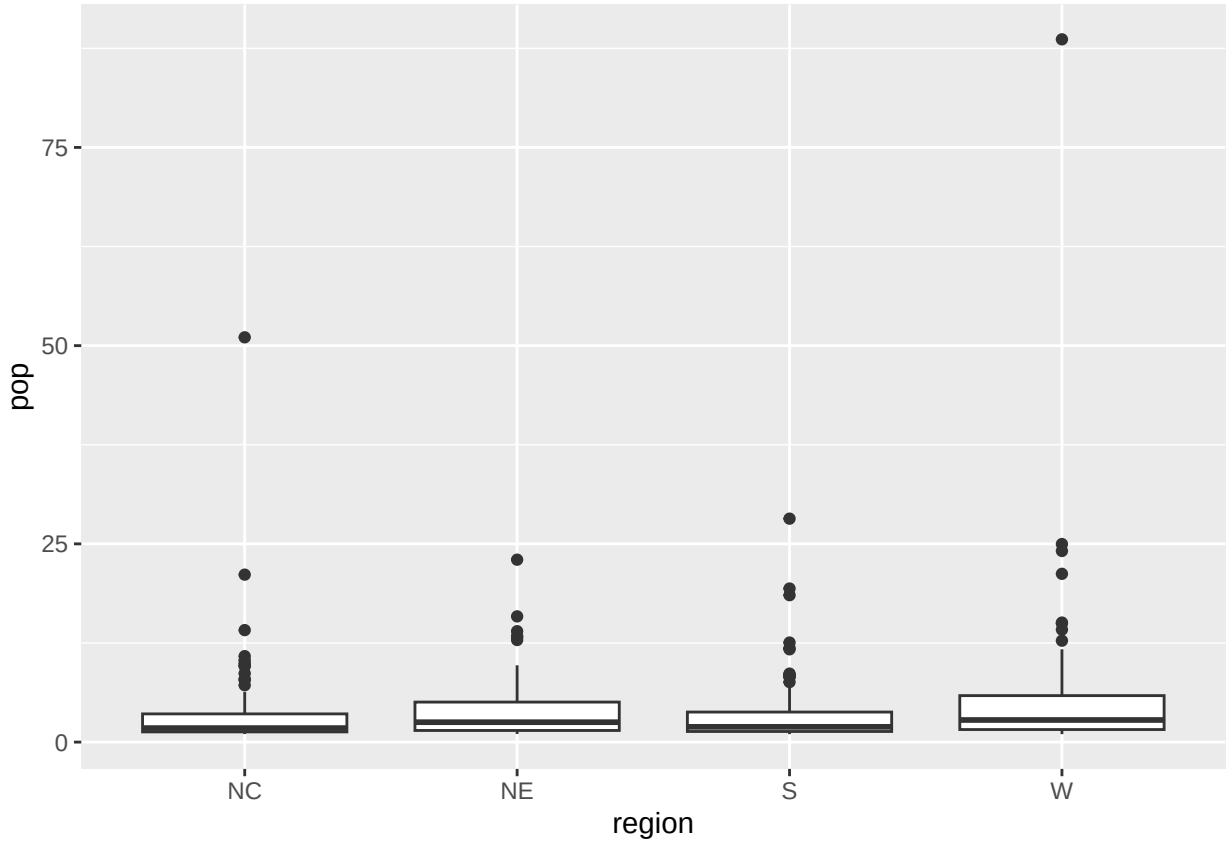
Use an appropriate plot to explore the association between population and number of physicians.

```
ggplot(cdi, aes(x = physicians, y = pop)) +
  geom_point()
```



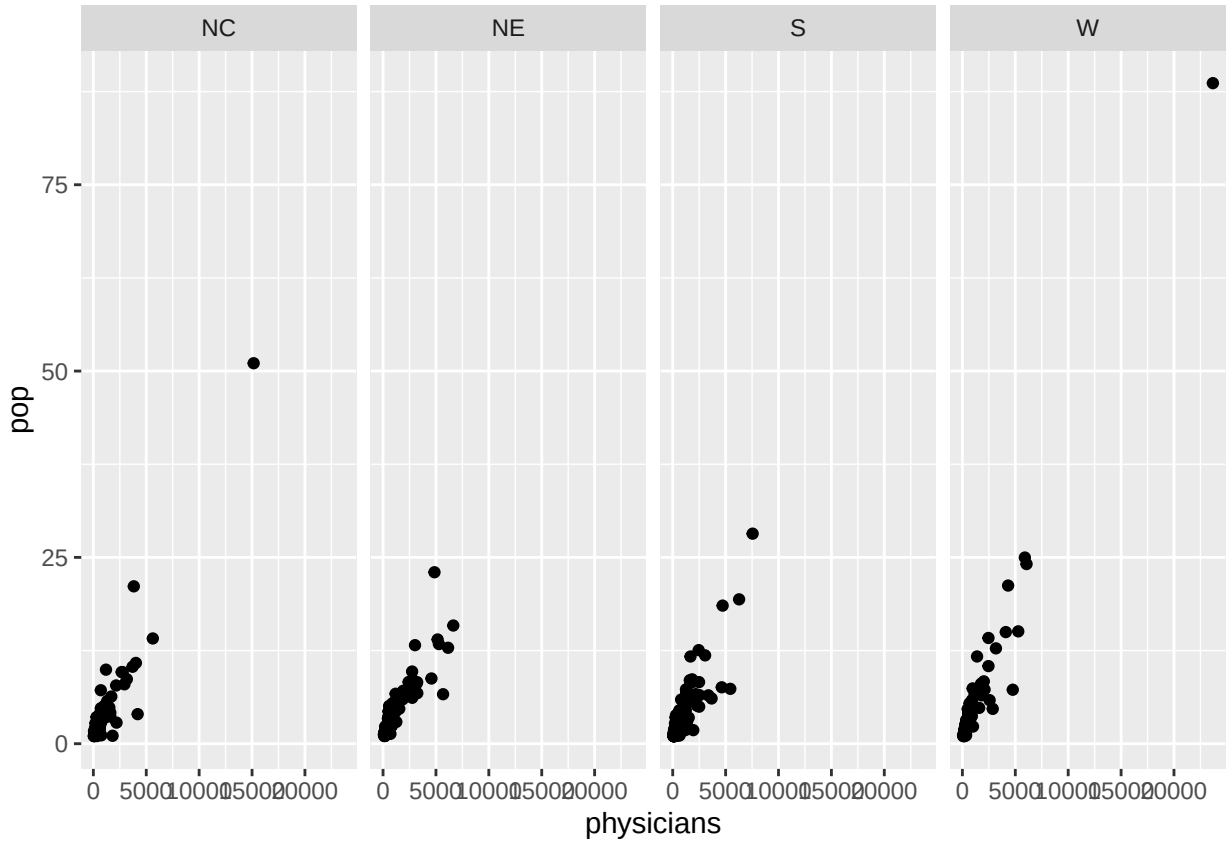
Use an appropriate plot to explore the association between region and population

```
ggplot(cdi, aes(x = region, y = pop)) +
  geom_boxplot()
```



Use an appropriate plot to explore the association between population and number of physicians in each region.

```
ggplot(cdi, aes(y = pop, x = physicians)) +
  geom_point() +
  facet_grid(~ region)
```



Create four datasets, one for each region.

```
cdi_NC <- filter(cdi, region == "NC")
head(cdi_NC)
```

```
## # A tibble: 6 x 17
##      id county    state  area   pop percent_18_34 percent_65 physicians  beds
##   <dbl> <chr>     <chr> <dbl> <dbl>         <dbl>      <dbl>      <dbl> <dbl>
## 1     2 Cook      IL      946 51.1           29.2       12.4      15153 21550
## 2     8 Wayne     MI      614 21.1           27.4       12.5       3823  9490
## 3    15 Cuyahoga  OH      458 14.1           26.3       15.6       5620  8132
## 4    25 Oakland   MI      873 10.8           27.6       10.9       4020  3254
## 5    27 Hennepin  MN      557 10.3           31.6       11.3       3706  5395
## 6    28 St._Louis MO      508  9.94          26.1       13.1       1194  1056
## # i 8 more variables: crimes <dbl>, high_school <dbl>, bachelors <dbl>,
## #   poverty <dbl>, unemployment <dbl>, capita_income <dbl>, total_income <dbl>,
## #   region <chr>
```

```
cdi_NE <- filter(cdi, region == "NE")
head(cdi_NE)
```

```
## # A tibble: 6 x 17
##      id county       state  area   pop percent_18_34 percent_65 physicians  beds
##   <dbl> <chr>        <chr> <dbl> <dbl>         <dbl>      <dbl>      <dbl> <dbl>
## 1     6 Kings        NY       71  23.0          28.3       12.4       4861  8942
```

```
## 2    11 Philadelphia PA      135  15.9          29.1       15.2       6641 10494
## 3    16 Middlesex    MA      824  14.0          31.7       12.5       5158  4152
## 4    17 Allegheny    PA      730  13.4          26.2       17.4       5281  8436
## 5    18 Suffolk      NY      911  13.2          27.9       10.8       3021  3904
## 6    19 Nassau       NY      287  12.9          25.7       14.2       6147  5200
## # i 8 more variables: crimes <dbl>, high_school <dbl>, bachelors <dbl>,
## #   poverty <dbl>, unemployment <dbl>, capita_income <dbl>, total_income <dbl>,
## #   region <chr>
```

```
cdi_S <- filter(cdi, region == "S")
head(cdi_S)
```

```
## # A tibble: 6 x 17
##      id county  state  area   pop percent_18_34 percent_65 physicians  beds
##   <dbl> <chr>   <chr> <dbl> <dbl>         <dbl>      <dbl>      <dbl> <dbl>
## 1     3 Harris  TX     1729  28.2          31.3        7.1       7553 12449
## 2     9 Dade    FL     1945  19.4          27.1       13.9       6274  8840
## 3    10 Dallas  TX      880  18.5          32.6        8.2       4718  6934
## 4    21 Broward FL     1209  12.6          25.3       20.7       2456  5543
## 5    22 Bexar   TX     1247  11.9          29.5        9.9       3062  4086
## 6    24 Tarrant TX      864  11.7          32.2        8.3       1677  3672
## # i 8 more variables: crimes <dbl>, high_school <dbl>, bachelors <dbl>,
## #   poverty <dbl>, unemployment <dbl>, capita_income <dbl>, total_income <dbl>,
## #   region <chr>
```

```
cdi_W <- filter(cdi, region == "W")
head(cdi_W)
```

```
## # A tibble: 6 x 17
##      id county      state  area   pop percent_18_34 percent_65 physicians  beds
##   <dbl> <chr>       <chr> <dbl> <dbl>         <dbl>      <dbl>      <dbl> <dbl>
## 1     1 Los_Angeles CA     4060  88.6          32.1        9.7      23677 27700
## 2     4 San_Diego   CA     4205  25.0          33.5       10.9       5905  6179
## 3     5 Orange      CA      790  24.1          32.6        9.2       6062  6369
## 4     7 Maricopa    AZ     9204  21.2          29.2       12.5       4320  6104
## 5    12 King        WA     2126  15.1          30.1       11.1       5280  4009
## 6    13 Santa_Clara CA     1291  15.0          32.6        8.7       4101  3342
## # i 8 more variables: crimes <dbl>, high_school <dbl>, bachelors <dbl>,
## #   poverty <dbl>, unemployment <dbl>, capita_income <dbl>, total_income <dbl>,
## #   region <chr>
```

Reproduce the following using markdown:

1. It was the best of times,
2. It was the worst of times,

- **It was the age of wisdom**,
- *it was the age of foolishness*,
- it was the epoch of belief...